ELSEVIER

contributions to economic analysis

Badi H. Baltagi
Editor

Panel Data Econometrics

Theoretical Contributions and Empirical Applications

PANEL DATA ECONOMETRICS
Theoretical Contributions and Empirical Applications

CONTRIBUTIONS
TO
ECONOMIC ANALYSIS

274

ELSEVIER

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo

PANEL DATA ECONOMETRICS

Theoretical Contributions and Empirical Applications

Edited by

BADI H. BALTAGI
Department of Economics and Center for Policy Research
Syracuse University, Syracuse, NY 13244-1020
U.S.A.

ELSEVIER

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo

ELSEVIER
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2006

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN-13: 978-0-444-52172-9
ISBN-10: 0-444-52172-0

For information on all Elsevier publictions
visit our website at books.elsevier.com

Printed and bound in The Netherlands

06 07 08 09 10 10 9 8 7 6 5 4 3 2 1

Introduction to the Series

This series consists of a number of hitherto unpublished studies, which are introduced by the editors in the belief that they represent fresh contributions to economic science.

The term ‘economic analysis’ as used in the title of the series has been adopted because it covers both the activities of the theoretical economist and the research worker.

Although the analytical method used by the various contributors are not the same, they are nevertheless conditioned by the common origin of their studies, namely theoretical problems encountered in practical research. Since for this reason, business cycle research and national accounting, research work on behalf of economic policy, and problems of planning are the main sources of the subjects dealt with, they necessarily determine the manner of approach adopted by the authors. Their methods tend to be ‘practical’ in the sense of not being too far remote from application to actual economic conditions. In addition, they are quantitative.

It is the hope of the editors that the publication of these studies will help to stimulate the exchange of scientific information and to reinforce international cooperation in the field of economics.

The Editors

This page intentionally left blank

Contents

Preface

Panel data econometrics has evolved rapidly over the last decade. Dynamic panel data estimation, non-linear panel data methods and the phenomenal growth in non-stationary panel data econometrics makes this an exciting area of research in econometrics. The 11th international conference on panel data held at Texas A&M University, College Station, Texas, June 2004, witnessed about 150 participants and 100 papers on panel data.

This volume includes some of the papers presented at that conference and other solicited papers that made it through the refereeing process. *Theoretical econometrics contributions include*: Bai and Kao who suggest a factor model approach to model cross-section dependence in the panel co-integrated regression setting; Lejeune who proposes new estimation methods and some diagnostics tests for a general heteroskedastic error component model with unbalanced panel data; Ullah and Huang who study the finite sample properties of feasible GLS for the random effects model with non-normal errors; Kazemi and Crouchley who suggest a pragmatic approach to the problem of estimating a dynamic panel regression with random effects under various assumptions about the nature of the initial conditions; Krishnakumar who uses a generalized version of the Frisch–Waugh theorem to extend Mundlak's (1978) results for the error component model. *Empirical applications include*: Sickles and Williams who estimate a dynamic model of crime using panel data from the 1958 Philadelphia Birth Cohort study; Baltagi and Griffin who find that at least 4 structural breaks in a panel data on liquor consumption for 21 Swedish counties over the period 1956–1999; Boumahdi, Chaaban and Thomas who estimate a flexible AIDS demand model for agricultural imports into Lebanon incorporating a three-way error component model that allows for product, country and time effects as separate unobserved determinants of import demand; Biørn, Skjerpen and Wangen who are concerned with the analysis of heterogeneous log-linear relationships (and specifically Cobb–Douglas production functions) at the firm-level and at the corresponding aggregate industry level. They use unbalanced panel data on firms from two Norwegian manufacturing industries over the period 1972–1993; Cermeño and Grier who apply a model that accounts for conditional heteroskedasticity and cross-sectional dependence to a panel of monthly inflation rates of the G7 over the period 1978.2–2003.9; Yasar, Nelson and Rejesus who use plant level panel data for Turkish manufacturing industries to analyze the relative importance of short-run versus long-run dynamics of the export-productivity relationship; Drine and Rault who focus on developing countries and analyze the long-run relationship between real exchange rate and some macroeconomic variables, via panel unit root and cointegration tests; Harris, Tang and Tseng who quantify the impact of employee turnover on productivity using an Australian business longitudinal survey over the period 1994/5 to 1997/8; Kaltchev who uses proprietary and confidential panel data on 113 public U.S. companies over the period 1997–2003 to analyze the demand for Directors' and Officers' liability insurance; Ortega-Díaz who assesses how income inequality influences economic growth across 32 Mexican States over the period 1960–2002.

Theoretical econometrics contributions

Bai and Kao suggest a factor model approach to model cross-section dependence in the panel co-integrated regression setting. Factor models are used to study world business cycles as well as common macro shocks like international financial crises or oil price shocks. Factor models offer a significant reduction in the number of sources of cross-sectional dependence in panel data and they allow for heterogeneous response to common shocks through heterogeneous factor loadings. Bai and Kao suggest a continuous-updated fully modified estimator for this model and show that it has better finite sample performance than OLS and a two step fully modified estimator.

Lejeune proposes new estimation methods for a general heteroskedastic error component model with unbalanced panel data, namely the Gaussian pseudo maximum likelihood of order 2. In addition, Lejeune suggests some diagnostics tests for heteroskedasticity, misspecification testing using m-tests, Hausman type and Information type tests. Lejeune applies these methods to estimate and test a translog production function using an unbalanced panel of 824 French firms observed over the period 1979–1988.

Ullah and Huang study the finite sample properties of feasible GLS for the random effects model with non-normal errors. They study the effects of skewness and excess kurtosis on the bias and mean squared error of the estimator using asymptotic expansions. This is done for large N and fixed T, under the assumption that the first four moments of the error are finite.

Kazemi and Crouchley suggest a pragmatic approach to the problem of estimating a dynamic panel regression with random effects under various assumptions about the nature of the initial conditions. They find that the full maximum likelihood improves the consistency results if the relationships between random effects, initial conditions and explanatory variables are correctly specified. They illustrate this by testing a variety of different hypothetical models in empirical contexts. They use information criteria to select the best approximating model.

Krishnakumar uses a generalized version of the Frisch–Waugh theorem to extend Mundlak's (1978) results for the error component model with individual effects that are correlated with the explanatory variables. In particular, this extension is concerned with the presence of time invariant variables and correlated specific effects.

Empirical contributions

The paper by Sickles and Williams estimates a dynamic model of crime using panel data from the 1958 Philadelphia Birth Cohort study. Agents are rational and anticipate the future consequence of their actions. The authors investigate the role of social capital through the influence of social norms on the decision to participate in crime. They find that the initial level of social capital stock is important in determining the pattern of criminal involvement in adulthood.

The paper by Baltagi and Griffin uses panel data on liquor consumption for 21 Swedish counties over the period 1956–1999. It finds that at least 4 structural breaks are necessary to account for the sharp decline in per-capita liquor consumption over this period. The first structural break coincides with the 1980 advertising ban, but subsequent breaks do not appear linked to particular policy initiatives. Baltagi and Griffin interpret these results as taste change accounting for increasing concerns with health issues and changing drinking mores.

The paper by Boumahdi, Chaaban and Thomas estimate a flexible AIDS demand model for agricultural imports into Lebanon incorporating a three-way error component model that allows for product, country and time effects as separate unobserved determinants of import demand. In their application to trade in agricultural commodities the authors are primarily concerned with the estimation of import demand elasticities. Conventionally, such estimates are frequently obtained from time series data that ignore the substitution elasticities across commodities, and thus implicitly ignore the cross-sectional dimension of the data. Exhaustive daily transactions (both imports and exports) data are obtained from the Lebanese customs administration for the years 1997–2002. Restricting their attention to major agricultural commodities (meat, dairy products, cereals, animals and vegetable fats and sugar), they estimate an import share equation for European products as a function of own-price and competitors prices. Competition is taking place between European countries, Arab and regional countries, North and South America and the rest of the world. The import share equations are estimated by allowing for parameter heterogeneity across the 5 commodity groups, and tests for the validity of the multi-way error components specification are performed using unbalanced panel data. Estimation results show that this specification is generally supported by the data.

The paper by Biørn, Skjerpen and Wangen is concerned with the analysis of heterogeneous log-linear relationships (and specifically Cobb–Douglas production functions) at the firm-level and at the corresponding aggregate industry level. While the presence of aggregation bias in log-linear models is widely recognized, considerable empirical analysis continues to be conducted ignoring the problem. This paper derives a decomposition that highlights the source of biases that arise in aggregate work. It defines some aggregate elasticity measures and illustrates these in an empirical exercise based on firm-level data in two Norwegian manufacturing industries: The pulp and paper industry (2823 observations, 237 firms) and the basic metals industry (2078 observations, 166 firms) observed over the period 1972–1993.

The paper by Cermeño and Grier specify a model that accounts for conditional heteroskedasticity and cross-sectional dependence within a typical panel data framework. The paper applies this model to a panel of monthly inflation rates of the G7 over the period 1978.2–2003.9 and finds significant and quite persistent patterns of volatility and cross-sectional dependence. The authors use the model to test two hypotheses about the inter-relationship between inflation and inflation uncertainty, finding no support for the hypothesis that higher inflation uncertainty produces higher average inflation rates and strong support for the hypothesis that higher inflation is less predictable.

The paper by Yasar, Nelson and Rejesus uses plant level panel data for Turkish manufacturing industries to analyze the relative importance of short-run versus long-run dynamics of the export-productivity relationship. The adopted econometric approach is a panel data error correction model that is estimated by means of system GMM. The data consists of plants with more than 25 employees from two industries, the textile and apparel industry and the motor vehicles and parts industry, observed over the period 1987–1997. They find that "permanent productivity shocks generate larger long-run export level responses, as compared to long-run productivity responses from permanent export shocks". This result suggests that industrial policy should be geared toward permanent improvements in plant-productivity in order to have sustainable long-run export and economic growth.

The paper by Drine and Rault focuses on developing countries and analyzes the long-run relationship between real exchange rate and some macroeconomic variables, via panel unit root and cointegration tests. The results show that the degrees of development and of openness of the economy strongly influence the real exchange rate. The panels considered are relatively small: Asia ($N = 7$, $T = 21$), Africa ($N = 21$, $T = 16$) and Latin America ($N = 17$, $T = 23$).

The paper by Harris, Tang and Tseng consider a balanced panel of medium sized firms drawn from the Australian business longitudinal survey over the period 1994/5 to 1997/8. The paper sets out to quantify the impact of employee turnover on productivity and finds that the optimal turnover rate is 0.22. This is higher than the sample median of 0.14 which raises the question about whether there are institutional rigidities hindering resource allocation in the labor market.

The paper by Kaltchev uses proprietary and confidential panel data on 113 public U.S. companies over the period 1997–2003 to analyze the demand for Directors' and Officers' liability insurance. Applying system GMM methods to a dynamic panel data model on this insurance data, Kaltchev rejects that this theory is habit driven but still finds some role for persistence. He also confirms the hypothesis that smaller companies demand more insurance. Other empirical findings include the following: Returns are significant in determining the amount of insurance and companies in financial distress demand higher insurance limits. Indicators of financial health such as leverage and volatility are significant, but not corporate governance.

The paper by Ortega-Díaz assesses how income inequality influences economic growth across 32 Mexican States over the period 1960–2002. Using dynamic panel data analysis, with both, urban personal income for grouped data and household income from national surveys, Ortega-Díaz finds that inequality and growth are positively related. This relationship is stable across variable definitions and data sets, but varies across regions and trade periods. A negative influence of inequality on growth is found in a period of restrictive trade policies. In contrast, a positive relationship is found in a period of trade openness.

I hope the readers enjoy this set of 15 papers on panel data and share my view on the wide spread use of panels in all fields of economics as clear from the applications. I would like to thank the anonymous referees that helped in reviewing these manuscripts. Also, Jennifer Broaddus for her editorial assistance and handling of these manuscripts.

Badi H. Baltagi
College Station, Texas and Syracuse, New York

This page intentionally left blank

List of Contributors

Numbers in parenthesis indicate the pages where the authors' contributions can be found.

Jushan Bai (3) Department of Economics, New York University, New York, NY 10003, USA and Department of Economics, Tsinghua University, Beijing 10084, China. E-mail: jushan.bai@nyu.edu

Badi H. Baltagi (167) Department of Economics, and Center for Policy Research, Syracuse University, Syracuse, NY 13244-1020, USA.
E-mail: bbaltagi@maxwell.syr.edu

Erik Biørn (229) Department of Economics, University of Oslo, 0317 Oslo, Norway and Research Department, Statistics Norway, 0033 Oslo, Norway.
E-mail: erik.biorn@econ.uio.no

Rachid Boumahdi (193) University of Toulouse, GREMAQ and LIHRE, F31000 Toulouse, France. E-mail: rachid.boumahdi@univ-tlse1.fr

Rodolfo Cermeño (259) División de Economía, CIDE, México D.F., México.
E-mail: rodolfo.cermeno@cide.edu

Jad Chaaban (193) University of Toulouse, INRA-ESR, F-31000 Toulouse cedex, France. E-mail: chaaban@toulouse.inra.fr

Rob Crouchley (91) Centre for e-Science, Fylde College, Lancaster University, Lancaster LA1 4YF, UK. E-mail: r.crouchley@lancaster.ac.uk

Imed Drine (307) Paris I, Masion des Sciences de l'Economie, 75647 Paris cedex 13, France. E-mail: drine@univ-paris1.fr

Kevin B. Grier (259) Department of Economics, University of Oklahoma, OK 73019, USA. E-mail: angus@ou.edu

James M. Griffin (167) Bush School of Government and Public Service, Texas A&M University, College Station, TX 77843-4220, USA.
E-mail: jgriffin@bushschool.tamu.edu

Mark N. Harris (327) Department of Econometrics and Business Statistics, Monash University, Melbourne, Vic 3800, Australia.
E-mail: mark.harris@buseco.monash.edu.au

Xiao Huang (67) Department of Economics University of California, Riverside, CA 92521-0427, USA. E-mail: xiao.huang@email.ucr.edu

George D. Kaltchev (351) Department of Economics, Southern Methodist University, Dallas, TX 75275-0496, USA. E-mail: gkaltche@mail.smu.edu

Chihwa Kao (3) Center for Policy Research and Department of Economics, Syracuse University Syracuse, NY 13244-1020, USA.
E-mail: cdkao@maxwell.syr.edu

Iraj Kazemi (91) Centre for Applied Statistics, Lancaster University, Lancaster LA1 4YF, UK. E-mail: i.kazemi@lancaster.ac.uk

Jaya Krishnakumar (119) Department of Econometrics, University of Geneva, UNI-MAIL, CH-1211 Geneva 4, Switzerland. E-mail: jaya.krishnakumar@metri.unige.ch

Bernard Lejeune (31) HEC-University of Liège, CORE and ERUDITE, 4000 Liège, Belgium. E-mail: b.lejeune@ulg.ac.be

Carl H. Nelson (279) Department of Agricultural & Consumer Economics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. E-mail: chnelson@uiuc.edu

Araceli Ortega-Díaz (361) Tecnológico de Monterrey, 14380 Tlalpan, México. E-mail: araceli.ortega@itesm.mx;aortega@sedesal.gob.mx

Chrisophe Rault (307) University of Evry-Val d'Essonne, Department d'économie, 91025 Evry cedex, France. E-mail: chrault@hotmail.com

Roderick M. Rejesus (279) Department of Agricultural & Applied Economics, Texas Tech University, Lubbock, TX 79409-2132, USA. E-mail: roderick.rejesus@ttu.edu

Robin C. Sickles (135) Department of Economics, Rice University, Houston, TX 77005-1892, USA. E-mail: rsickles@rice.edu

Terje Skjerpen (229) Research Department, Statistics Norway, 0033 Oslo, Norway. E-mail: terje.skjerpen@ssb.no

Kam-Ki Tang (327) School of Economics, University of Queensland, St. Lucia, Qld 4072, Australia. E-mail: kk.tang@uq.edu.au

Alban Thomas (193) University of Toulouse, INRA-LERNA, F-31000 Toulouse cedex, France. E-mail: thomas@toulouse.inra.fr

Yi-Ping Tseng (327) Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Parkville, Vic 3010, Australia. E-mail: y.tseng@unimelb.edu.au

Aman Ullah (67) Department of Economics, University of California, Riverside, CA 92521-0427, USA. E-mail: aman.ullah@ucr.edu

Knut R. Wangen (229) Research Department, Statistics Norway, 0033 Oslo, Norway. E-mail: knut.reidar.wangen@ssb.no

Jenny Williams (135) Department of Economics, University of Melbourne, Melbourne, Vic 3010, Australia. E-mail: jenny.williams@unimelb.edu.au

Mahmut Yasar (279) Department of Economics, Emory University, Atlanta, GA 30322, USA. E-mail: myasar@emory.edu

PART I

Theoretical Contributions

This page intentionally left blank

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74001-9

CHAPTER 1

On the Estimation and Inference of a Panel Cointegration Model with Cross-Sectional Dependence

Jushan Bai[a] and Chihwa Kao[b]

[a]Department of Economics, New York University, New York, NY 10003, USA and Department of Economics, Tsinghua University, Beijing 10084, China
E-mail address: Jushan.Bai@nyu.edu
[b]Center for Policy Research and Department of Economics, Syracuse University, Syracuse, NY 13244-1020, USA
E-mail address: cdkao@maxwell.syr.edu

Abstract

Most of the existing literature on panel data cointegration assumes cross-sectional independence, an assumption that is difficult to satisfy. This paper studies panel cointegration under cross-sectional dependence, which is characterized by a factor structure. We derive the limiting distribution of a fully modified estimator for the panel cointegrating coefficients. We also propose a continuous-updated fully modified (CUP-FM) estimator. Monte Carlo results show that the CUP-FM estimator has better small sample properties than the two-step FM (2S-FM) and OLS estimators.

Keywords: panel data, cross-sectional dependence, factor analysis, CUP-FM, 2S-FM

JEL classifications: C13, C33

1.1 Introduction

A convenient but difficult to justify assumption in panel cointegration analysis is cross-sectional independence. Left untreated, cross-sectional dependence causes bias and inconsistency estimation, as argued by Andrews (2005). In this paper, we use a factor structure to characterize cross-sectional dependence. Factors models are especially suited for this purpose. One major source of cross-section correlation in macroeconomic data is common shocks, e.g., oil price shocks and international financial crises. Common shocks drive the underlying comovement of economic variables. Factor models provide an effective way to extract the comovement and have been used in various studies.[1] Cross-sectional correlation exists even in micro level data because of herd behavior (fashions, fads, and imitation cascades) either at firm level or household level. The general state of an economy (recessions or booms) also affects household decision making. Factor models accommodate individual's different responses to common shocks through heterogeneous factor loadings.

Panel data models with correlated cross-sectional units are important due to increasing availability of large panel data sets and increasing interconnectedness of the economies. Despite the immense interest in testing for panel unit roots and cointegration,[2] not much attention has been paid to the issues of cross-sectional dependence. Studies using factor models for nonstationary data include Bai and Ng (2004), Bai (2004), Phillips and Sul (2003), and Moon and Perron (2004). Chang (2002) proposed to use a nonlinear IV estimation to construct a new panel unit root test. Hall *et al.* (1999) considered a problem of determining the number of common trends. Baltagi *et al.* (2004) derived several Lagrange Multiplier tests for the panel data regression model with spatial error correlation. Robertson and Symon (2000), Coakley *et al.* (2002) and Pesaran (2004) proposed to use common factors to capture the cross-sectional dependence in stationary panel models. All these studies focus on either stationary data or panel unit root studies rather than panel cointegration.

This paper makes three contributions. First, it adds to the literature by suggesting a factor model for panel cointegrations. Second, it proposes a continuous-updated fully modified (CUP-FM) estimator. Third, it provides a comparison for the finite sample properties of the OLS, two-step fully modified (2S-FM), CUP-FM estimators.

The rest of the paper is organized as follows. Section 1.2 introduces the model. Section 1.3 presents assumptions. Sections 1.4 and 1.5 develop the asymptotic theory for the OLS and fully modified (FM) estimators. Section 1.6 discusses a feasible FM estimator and suggests a CUP-FM estimator. Section 1.7 makes some remarks on hypothesis testing. Section 1.8 presents Monte Carlo results to illustrate the finite sample properties of the OLS and FM estimators. Section 1.9 summarizes the findings. Appendix A1 contains the proofs of lemmas and theorems.

The following notations are used in the paper. We write the integral $\int_0^1 W(s)\,\mathrm{d}s$ as $\int W$ when there is no ambiguity over limits. We define

[1] For example, Stock and Watson (2002), Gregory and Head (1999), Forni and Reichlin (1998) and Forni *et al.* (2000) to name a few.

[2] See Baltagi and Kao (2000) for a recent survey.

$\Omega^{1/2}$ to be any matrix such that $\Omega = (\Omega^{1/2})(\Omega^{1/2})'$. We use $\|A\|$ to denote $\{\mathrm{tr}(A'A)\}^{1/2}$, $|A|$ to denote the determinant of A, $\Rightarrow$ to denote weak convergence, $\xrightarrow{p}$ to denote convergence in probability, $[x]$ to denote the largest integer $\leqslant x$, $I(0)$ and $I(1)$ to signify a time-series that is integrated of order zero and one, respectively, and $BM(\Omega)$ to denote Brownian motion with the covariance matrix Ω. We let $M < \infty$ be a generic positive number, not depending on T or n.

1.2 The model

Consider the following fixed effect panel regression:

$$y_{it} = \alpha_i + \beta x_{it} + e_{it}, \quad i = 1, \ldots, n, \ t = 1, \ldots, T, \tag{1.1}$$

where y_{it} is 1×1, β is a $1 \times k$ vector of the slope parameters, α_i is the intercept, and e_{it} is the stationary regression error. We assume that x_{it} is a $k \times 1$ integrated processes of order one for all i, where

$$x_{it} = x_{it-1} + \varepsilon_{it}.$$

Under these specifications, (1.1) describes a system of cointegrated regressions, i.e., y_{it} is cointegrated with x_{it}. The initialization of this system is $y_{i0} = x_{i0} = \mathrm{O}_p(1)$ as $T \to \infty$ for all i. The individual constant term α_i can be extended into general deterministic time trends such as $\alpha_{0i} + \alpha_{1i}t + \cdots + \alpha_{pi}t$ or other deterministic component. To model the cross-sectional dependence we assume the error term, e_{it}, follows a factor model (e.g., Bai and Ng, 2002, 2004):

$$e_{it} = \lambda_i' F_t + u_{it}, \tag{1.2}$$

where F_t is a $r \times 1$ vector of common factors, λ_i is a $r \times 1$ vector of factor loadings and u_{it} is the idiosyncratic component of e_{it}, which means

$$E(e_{it}e_{jt}) = \lambda_i' E(F_t F_t')\lambda_j,$$

i.e., e_{it} and e_{jt} are correlated due to the common factors F_t.

REMARK 1.1. We could also allow ε_{it} to have a factor structure such that

$$\varepsilon_{it} = \gamma_i' F_t + \eta_{it}.$$

Then we can use Δx_{it} to estimate F_t and γ_i. Or we can use e_{it} together with Δx_{it} to estimate F_t, λ_i and γ_i. In general, ε_{it} can be of the form

$$\varepsilon_{it} = \gamma_i' F_t + \tau_i' G_t + \eta_{it},$$

where F_t and G_t are zero mean processes, and η_{it} are usually independent over i and t.

1.3 Assumptions

Our analysis is based on the following assumptions.

ASSUMPTION 1.1. As $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n}\lambda_i\lambda_i' \to \Sigma_\lambda$, a $r \times r$ positive definite matrix.

ASSUMPTION 1.2. Let $w_{it} = (F_t', u_{it}, \varepsilon_{it}')'$. For each i, $w_{it} = \Pi_i(L)v_{it} = \sum_{j=0}^{\infty}\Pi_{ij}v_{it-j}$, $\sum_{j=0}^{\infty} j^a\|\Pi_{ij}\| < \infty$, $|\Pi_i(1)| \neq 0$, for some $a > 1$, where v_{it} is i.i.d. over t. In addition, $Ev_{it} = 0$, $E(v_{it}v_{it}') = \Sigma_v > 0$, and $E\|v_{it}\|^8 \leqslant M < \infty$.

ASSUMPTION 1.3. F_t and u_{it} are independent; u_{it} are independent across i.

Under Assumption 1.2, a multivariate invariance principle for w_{it} holds, i.e., the partial sum process $\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} w_{it}$ satisfies:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} w_{it} \Rightarrow B(\Omega_i) \quad \text{as } T \to \infty \text{ for all } i, \tag{1.3}$$

where

$$B_i = \begin{bmatrix} B_F \\ B_{ui} \\ B_{\varepsilon i} \end{bmatrix}.$$

The long-run covariance matrix of $\{w_{it}\}$ is given by

$$\begin{aligned} \Omega_i &= \sum_{j=-\infty}^{\infty} E(w_{i0}w_{ij}') \\ &= \Pi_i(1)\Sigma_v\Pi_i(1)' \\ &= \Sigma_i + \Gamma_i + \Gamma_i' \\ &= \begin{bmatrix} \Omega_{Fi} & \Omega_{Fui} & \Omega_{F\varepsilon i} \\ \Omega_{uFi} & \Omega_{ui} & \Omega_{u\varepsilon i} \\ \Omega_{\varepsilon Fi} & \Omega_{\varepsilon ui} & \Omega_{\varepsilon i} \end{bmatrix}, \end{aligned}$$

where

$$\Gamma_i = \sum_{j=1}^{\infty} E(w_{i0}w_{ij}') = \begin{bmatrix} \Gamma_{Fi} & \Gamma_{Fui} & \Gamma_{F\varepsilon i} \\ \Gamma_{uFi} & \Gamma_{ui} & \Gamma_{u\varepsilon i} \\ \Gamma_{\varepsilon Fi} & \Gamma_{\varepsilon ui} & \Gamma_{\varepsilon i} \end{bmatrix} \tag{1.4}$$

and

$$\Sigma_i = E(w_{i0}w_{i0}') = \begin{bmatrix} \Sigma_{Fi} & \Sigma_{Fui} & \Sigma_{F\varepsilon i} \\ \Sigma_{uFi} & \Sigma_{ui} & \Sigma_{u\varepsilon i} \\ \Sigma_{\varepsilon Fi} & \Sigma_{\varepsilon ui} & \Sigma_{\varepsilon i} \end{bmatrix}$$

are partitioned conformably with w_{it}. We denote

$$\Omega = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Omega_i,$$

$$\Gamma = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Gamma_i,$$

and

$$\Sigma = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Sigma_i.$$

ASSUMPTION 1.4. $\Omega_{\varepsilon i}$ is nonsingular, i.e., $\{x_{it}\}$, are not cointegrated.

Define

$$\Omega_{bi} = \begin{bmatrix} \Omega_{Fi} & \Omega_{Fui} \\ \Omega_{uFi} & \Omega_{ui} \end{bmatrix}, \qquad \Omega_{b\varepsilon i} = \begin{bmatrix} \Omega_{F\varepsilon i} \\ \Omega_{u\varepsilon i} \end{bmatrix}$$

and

$$\Omega_{b.\varepsilon i} = \Omega_{bi} - \Omega_{b\varepsilon i}\Omega_{\varepsilon i}^{-1}\Omega_{\varepsilon bi}.$$

Then, B_i can be rewritten as

$$B_i = \begin{bmatrix} B_{bi} \\ B_{\varepsilon i} \end{bmatrix} = \begin{bmatrix} \Omega_{b.\varepsilon i}^{1/2} & \Omega_{b\varepsilon i}\Omega_{\varepsilon i}^{-1/2} \\ 0 & \Omega_{\varepsilon i}^{1/2} \end{bmatrix} \begin{bmatrix} V_{bi} \\ W_i \end{bmatrix}, \tag{1.5}$$

where

$$B_{bi} = \begin{bmatrix} B_F \\ B_{ui} \end{bmatrix},$$

$$V_{bi} = \begin{bmatrix} V_F \\ V_{ui} \end{bmatrix},$$

and

$$\begin{bmatrix} V_{bi} \\ W_i \end{bmatrix} = BM(I)$$

is a standardized Brownian motion. Define the one-sided long-run covariance

$$\begin{aligned}\varDelta_i &= \varSigma_i + \varGamma_i \\ &= \sum_{j=0}^{\infty} E(w_{i0}w_{ij}')\end{aligned}$$

with

$$\varDelta_i = \begin{bmatrix} \varDelta_{bi} & \varDelta_{b\varepsilon i} \\ \varDelta_{\varepsilon bi} & \varDelta_{\varepsilon i} \end{bmatrix}.$$

REMARK 1.2. (1) Assumption 1.1 is a standard assumption in factor models (e.g., Bai and Ng, 2002, 2004) to ensure the factor structure is identifiable. We only consider nonrandom factor loadings for simplicity. Our results still hold when the λ_i's are random, provided they are independent of the factors and idiosyncratic errors, and $E\|\lambda_i\|^4 \leqslant M$.

(2) Assumption 1.2 assumes that the random factors, F_t, and idiosyncratic shocks $(u_{it}, \varepsilon_{it}')$ are stationary linear processes. Note that F_t and ε_{it} are allowed to be correlated. In particular, ε_{it} may have a factor structure as in Remark 1.1.

(3) Assumption of independence made in Assumption 1.3 between F_t and u_{it} can be relaxed following Bai and Ng (2002). Nevertheless, independence is not a restricted assumption since cross-sectional correlations in the regression errors e_{it} are taken into account by the common factors.

1.4 OLS

Let us first study the limiting distribution of the OLS estimator for Equation (1.1). The OLS estimator of β is

$$\hat{\beta}_{\mathrm{OLS}} = \left[\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}(x_{it}-\bar{x}_i)'\right]\left[\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\right]^{-1}. \tag{1.6}$$

THEOREM 1.1. *Under Assumptions 1.1–1.4, we have*

$$\begin{aligned}&\sqrt{n}T(\hat{\beta}_{\mathrm{OLS}}-\beta) - \sqrt{n}\delta_{nT} \\ &\quad \Rightarrow N\left(0, 6\varOmega_{\varepsilon}^{-1}\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\lambda_i'\varOmega_{F.\varepsilon i}\lambda_i\varOmega_{\varepsilon i} + \varOmega_{u.\varepsilon i}\varOmega_{\varepsilon i})\right\}\varOmega_{\varepsilon}^{-1}\right),\end{aligned}$$

as $(n, T \to \infty)$ *with* $\frac{n}{T} \to 0$ *where*

$$\delta_{nT} = \frac{1}{n}\left[\sum_{i=1}^{n} \lambda_i' \left(\Omega_{F\varepsilon i} \Omega_{\varepsilon i}^{1/2} \left(\frac{1}{T} \sum_{t=1}^{T} x_{it}'(x_{it} - \bar{x}_i) \right) \Omega_{\varepsilon i}^{-1/2} + \Delta_{F\varepsilon i} \right) \right.$$
$$\left. + \Omega_{u\varepsilon i} \Omega_{\varepsilon i}^{1/2} \left(\frac{1}{T} \sum_{t=1}^{T} x_{it}'(x_{it} - \bar{x}_i) \right) \Omega_{\varepsilon i}^{-1/2} + \Delta_{u\varepsilon i} \right]$$
$$\times \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{T^2} \sum_{t=1}^{T} (x_{it} - x_{it})(x_{it} - \bar{x}_i)' \right]^{-1},$$

$\widetilde{W}_i = W_i - \int W_i$ *and* $\Omega_\varepsilon = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Omega_{\varepsilon i}$.

REMARK 1.3. It is also possible to construct the bias-corrected OLS by using the averages of the long run covariances. Note

$$\begin{aligned}
&E[\delta_{nT}] \\
&\quad \simeq \frac{1}{n}\left[\sum_{i=1}^{n} \lambda_i' \left(-\frac{1}{2}\Omega_{F\varepsilon i} + \Delta_{F\varepsilon i} \right) - \frac{1}{2}\Omega_{u\varepsilon i} + \Delta_{u\varepsilon i} \right] \left(\frac{1}{6}\Omega_\varepsilon \right)^{-1} \\
&\quad = \frac{1}{n}\left[\sum_{i=1}^{n} \left(-\frac{1}{2} \right) (\lambda_i' \Omega_{F\varepsilon i} + \Omega_{u\varepsilon i}) + \lambda_i' \Delta_{F\varepsilon i} + \Delta_{u\varepsilon i} \right] \left(\frac{1}{6}\Omega_\varepsilon \right)^{-1} \\
&\quad = \left(\frac{1}{n} \sum_{i=1}^{n} \left(-\frac{1}{2} \right) \lambda_i' \Omega_{F\varepsilon i} + \frac{1}{n} \sum_{i=1}^{n} \Omega_{u\varepsilon i} + \frac{1}{n} \sum_{i=1}^{n} \lambda_i' \Delta_{F\varepsilon i} \right. \\
&\qquad \left. + \frac{1}{n} \sum_{i=1}^{n} \Delta_{u\varepsilon i} \right) \left(\frac{1}{6}\Omega_\varepsilon \right)^{-1}.
\end{aligned}$$

It can be shown by a central limit theorem that

$$\sqrt{n}\big(\delta_{nT} - E[\delta_{nT}]\big) \Rightarrow N(0, B)$$

for some B. Therefore,

$$\begin{aligned}
&\sqrt{n}T(\hat{\beta}_{\text{OLS}} - \beta) - \sqrt{n}E[\delta_{nT}] \\
&\qquad = \sqrt{n}T(\hat{\beta}_{\text{OLS}} - \beta) - \sqrt{n}\delta_{nT} + \sqrt{n}\big(\delta_{nT} - E[\delta_{nT}]\big) \\
&\quad \Rightarrow N(0, A)
\end{aligned}$$

for some A.

1.5 FM estimator

Next we examine the limiting distribution of the FM estimator, $\hat{\beta}_{\mathrm{FM}}$. The FM estimator was suggested by Phillips and Hansen (1990) in a different context (nonpanel data). The FM estimator is constructed by making corrections for endogeneity and serial correlation to the OLS estimator $\hat{\beta}_{\mathrm{OLS}}$ in (1.6). The endogeneity correction is achieved by modifying the variable y_{it}, in (1.1) with the transformation

$$y_{it}^{+} = y_{it} - (\lambda_i' \Omega_{F\varepsilon i} + \Omega_{u\varepsilon i}) \Omega_{\varepsilon i}^{-1} \Delta x_{it}.$$

The serial correlation correction term has the form

$$\begin{aligned} \Delta_{b\varepsilon i}^{+} &= \Delta_{b\varepsilon i} - \Omega_{b\varepsilon i} \Omega_{\varepsilon i}^{-1} \Delta_{\varepsilon i} \\ &= \begin{bmatrix} \Delta_{F\varepsilon i}^{+} \\ \Delta_{u\varepsilon i}^{+} \end{bmatrix}. \end{aligned}$$

Therefore, the infeasible FM estimator is

$$\begin{aligned} \tilde{\beta}_{\mathrm{FM}} = &\left[\sum_{i=1}^{n} \left(\sum_{t=1}^{T} y_{it}^{+} (x_{it} - \bar{x}_i)' - T \left(\lambda_i' \Delta_{F\varepsilon i}^{+} + \Delta_{u\varepsilon i}^{+} \right) \right) \right] \\ &\times \left[\sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right]^{-1}. \end{aligned} \tag{1.7}$$

Now, we state the limiting distribution of $\tilde{\beta}_{\mathrm{FM}}$.

THEOREM 1.2. *Let Assumptions* 1.1–1.4 *hold. Then as* $(n, T \to \infty)$ *with* $\frac{n}{T} \to 0$

$$\begin{aligned} &\sqrt{n} T (\tilde{\beta}_{\mathrm{FM}} - \beta) \\ &\quad \Rightarrow N\left(0, 6\Omega_{\varepsilon}^{-1} \left\{ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (\lambda_i' \Omega_{F.\varepsilon i} \lambda_i \Omega_{\varepsilon i} + \Omega_{u.\varepsilon i} \Omega_{\varepsilon i}) \right\} \Omega_{\varepsilon}^{-1} \right). \end{aligned}$$

REMARK 1.4. The asymptotic distribution in Theorem 1.2 is reduced to

$$\sqrt{n} T (\tilde{\beta}_{\mathrm{FM}} - \beta) \Rightarrow \mathrm{N}\left(0, 6\Omega_{\varepsilon}^{-1} \left(\left(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \lambda_i^2 \right) \Omega_{F.\varepsilon} + \Omega_{u.\varepsilon} \right) \right)$$

if the long-run covariances are the same across the cross-sectional unit i and $r = 1$.

1.6 Feasible FM

In this section we investigate the limiting distribution of the feasible FM. We will show that the limiting distribution of the feasible FM is not affected when λ_i, $\Omega_{\varepsilon i}$, $\Omega_{\varepsilon bi}$, $\Omega_{\varepsilon i}$, and $\Delta_{\varepsilon bi}$ are estimated. To estimate λ_i, we use the method of principal components used in Stock and Watson (2002). Let $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$ and $F = (F_1, F_2, \ldots, F_T)'$. The method of principal components of λ and F minimizes

$$V(r) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (\hat{e}_{it} - \lambda_i' F_t)^2,$$

where

$$\begin{aligned} \hat{e}_{it} &= y_{it} - \hat{\alpha}_i - \hat{\beta} x_{it} \\ &= (y_{it} - \bar{y}_i) - \hat{\beta}(x_{it} - \bar{x}_i), \end{aligned}$$

with a consistent estimator $\hat{\beta}$. Concentrating out λ and using the normalization that $F'F/T = I_r$, the optimization problem is identical to maximizing $\operatorname{tr}(F'(ZZ')F)$, where $Z = (\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n)$ is $T \times n$ with $\hat{e}_i = (\hat{e}_{i1}, \hat{e}_{i2}, \ldots, \hat{e}_{iT})'$. The estimated factor matrix, denoted by $\widehat{F}$, a $T \times r$ matrix, is $\sqrt{T}$ times eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix ZZ', and

$$\begin{aligned} \hat{\lambda}' &= (\widehat{F}'\widehat{F})^{-1}\widehat{F}'Z \\ &= \frac{\widehat{F}'Z}{T} \end{aligned}$$

is the corresponding matrix of the estimated factor loadings. It is known that the solution to the above minimization problem is not unique, i.e., λ_i and F_t are not directly identifiable but they are identifiable up to a transformation H. For our setup, knowing $H\lambda_i$ is as good as knowing λ_i. For example in (1.7) using $\lambda_i' \Delta_{F\varepsilon i}^{+}$ will give the same information as using $\lambda_i' H' H'^{-1} \Delta_{F\varepsilon i}^{+}$ since $\Delta_{F\varepsilon i}^{+}$ is also identifiable up to a transformation, i.e., $\lambda_i' H' H'^{-1} \Delta_{F\varepsilon i}^{+} = \lambda_i' \Delta_{F\varepsilon i}^{+}$. Therefore, when assessing the properties of the estimates we only need to consider the differences in the space spanned by, say, between $\hat{\lambda}_i$ and λ_i.

Define the feasible FM, $\hat{\beta}_{\mathrm{FM}}$, with $\hat{\lambda}_i$, $\widehat{F_t}$, $\widehat{\Sigma_i}$, and $\widehat{\Omega_i}$ in place of λ_i, F_t, Σ_i, and Ω_i,

$$\begin{aligned} \hat{\beta}_{\mathrm{FM}} = &\left[\sum_{i=1}^{n} \left(\sum_{t=1}^{T} \hat{y}_{it}^{+} (x_{it} - \bar{x}_i)' - T\left(\hat{\lambda}_i' \widehat{\Delta}_{F\varepsilon i}^{+} + \widehat{\Delta}_{u\varepsilon i}^{+}\right) \right) \right] \\ &\times \left[\sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right]^{-1}, \end{aligned}$$

where

$$\hat{y}_{it}^{+} = y_{it} - (\hat{\lambda}_i' \widehat{\Omega}_{F\varepsilon i} + \widehat{\Omega}_{u\varepsilon i})\widehat{\Omega}_{\varepsilon i}^{-1} \Delta x_{it}$$

and $\widehat{\Delta}_{F\varepsilon i}^{+}$ and $\widehat{\Delta}_{u\varepsilon i}^{+}$ are defined similarly.

Assume that $\Omega_i = \Omega$ for all i. Let

$$e_{it}^{+} = e_{it} - (\lambda_i' \Omega_{F\varepsilon} + \Omega_{u\varepsilon})\Omega_{\varepsilon}^{-1} \Delta x_{it},$$

$$\widehat{\Delta}_{b\varepsilon n}^{+} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\Delta}_{b\varepsilon i}^{+},$$

and

$$\Delta_{b\varepsilon n}^{+} = \frac{1}{n} \sum_{i=1}^{n} \Delta_{b\varepsilon i}^{+}.$$

Then

$$\begin{aligned}
&\sqrt{n}T(\hat{\beta}_{\mathrm{FM}} - \tilde{\beta}_{\mathrm{FM}}) \\
&\quad = \frac{1}{\sqrt{n}T} \sum_{i=1}^{n} \left\{ \left(\sum_{t=1}^{T} \hat{e}_{it}^{+}(x_{it} - \bar{x}_i)' - T\widehat{\Delta}_{b\varepsilon n}^{+} \right) \right. \\
&\qquad \left. - \left(\sum_{t=1}^{T} e_{it}^{+}(x_{it} - \bar{x}_i)' - T\Delta_{b\varepsilon n}^{+} \right) \right\} \\
&\qquad \times \left[\frac{1}{nT^2} \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right]^{-1} \\
&\quad = \left[\frac{1}{\sqrt{n}T} \sum_{i=1}^{n} \left(\sum_{t=1}^{T} (\hat{e}_{it}^{+} - e_{it}^{+})(x_{it} - \bar{x}_i)' - T\big(\widehat{\Delta}_{b\varepsilon n}^{+} - \Delta_{b\varepsilon n}^{+}\big) \right) \right] \\
&\qquad \times \left[\frac{1}{nT^2} \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right]^{-1}.
\end{aligned}$$

Before we prove Theorem 1.3 we need the following lemmas.

LEMMA 1.1. *Under Assumptions* 1.1–1.4 $\sqrt{n}(\widehat{\Delta}_{b\varepsilon n}^{+} - \Delta_{b\varepsilon n}^{+}) = \mathrm{o}_p(1)$.

Lemma 1.1 can be proved similarly by following Phillips and Moon (1999) and Moon and Perron (2004).

LEMMA 1.2. *Suppose Assumptions* 1.1–1.4 *hold. There exists an H with rank r such that as* $(n, T \to \infty)$

(i) $$\frac{1}{n}\sum_{i=1}^{n}\|\hat{\lambda}_i - H\lambda_i\|^2 = \mathrm{O}_p\left(\frac{1}{\delta_{nT}^2}\right).$$

(ii) *Let* c_i $(i = 1, 2, \dots, n)$ *be a sequence of random matrices such that* $c_i = \mathrm{O}_p(1)$ *and* $\frac{1}{n}\sum_{i=1}^{n}\|c_i\|^2 = \mathrm{O}_p(1)$ *then*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\lambda}_i - H\lambda_i)'c_i = \mathrm{O}_p\left(\frac{1}{\delta_{nT}^2}\right),$$

where $\delta_{nT} = \min\{\sqrt{n}, \sqrt{T}\}$.

Bai and Ng (2002) showed that for a known $\hat{e}_{it}$ that the average squared deviations between $\hat{\lambda}_i$ and $H\lambda_i$ vanish as n and T both tend to infinity and the rate of convergence is the minimum of n and T. Lemma 1.2 can be proved similarly by following Bai and Ng (2002) that parameter estimation uncertainty for β has no impact on the null limit distribution of $\hat{\lambda}_i$.

LEMMA 1.3. *Under Assumptions 1.1–1.4*

$$\frac{1}{\sqrt{n}T}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\hat{e}_{it}^{+} - e_{it}^{+}\right)(x_{it} - \bar{x}_i)' = \mathrm{o}_p(1)$$

as $(n, T \to \infty)$ *and* $\frac{\sqrt{n}}{T} \to 0$.

Then we have the following theorem:

THEOREM 1.3. *Under Assumptions 1.1–1.4 and* $(n, T \to \infty)$ *and* $\frac{\sqrt{n}}{T} \to 0$

$$\sqrt{n}T(\hat{\beta}_{\mathrm{FM}} - \tilde{\beta}_{\mathrm{FM}}) = \mathrm{o}_p(1).$$

In the literature, the FM-type estimators usually were computed with a two-step procedure, by assuming an initial consistent estimate of β, say $\hat{\beta}_{\mathrm{OLS}}$. Then, one constructs estimates of the long-run covariance matrix, $\widehat{\Omega}^{(1)}$, and loading, $\hat{\lambda}_i^{(1)}$. The 2S-FM, denoted $\hat{\beta}_{2S}^{(1)}$ is obtained using $\widehat{\Omega}^{(1)}$ and $\hat{\lambda}_i^{(1)}$:

$$\hat{\beta}_{2S}^{(1)} = \left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\hat{y}_{it}^{+(1)}(x_{it} - \bar{x}_i)' - T\left(\hat{\lambda}_i'^{(1)}\widehat{\Delta}_{F\varepsilon i}^{+(1)} + \widehat{\Delta}_{u\varepsilon i}^{+(1)}\right)\right)\right] \times \left[\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'\right]^{-1}. \qquad (1.8)$$

In this paper, we propose a CUP-FM estimator. The CUP-FM is constructed by estimating parameters and long-run covariance matrix and loading recursively. Thus $\hat{\beta}_{\mathrm{FM}}$, $\widehat{\Omega}$ and $\hat{\lambda}_i$ are estimated repeatedly, until convergence is reached. In Section 1.8, we find the CUP-FM has a superior small sample properties as compared with the 2S-FM, i.e., CUP-FM has smaller bias than the common 2S-FM estimator. The CUP-FM is defined as

$$\begin{aligned}\hat{\beta}_{\mathrm{CUP}} = &\Bigg[\sum_{i=1}^{n}\Bigg(\sum_{t=1}^{T}\hat{y}_{it}^{+}(\hat{\beta}_{\mathrm{CUP}})(x_{it}-\bar{x}_i)' \\ &- T\big(\hat{\lambda}_i'(\hat{\beta}_{\mathrm{CUP}})\widehat{\Delta}_{F\varepsilon i}^{+}(\hat{\beta}_{\mathrm{CUP}}) + \widehat{\Delta}_{u\varepsilon i}^{+}(\hat{\beta}_{\mathrm{CUP}})\big)\Bigg)\Bigg] \\ &\times\Bigg[\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\Bigg]^{-1}.\end{aligned} \tag{1.9}$$

REMARK 1.5. (1) In this paper, we assume the number of factors, r, is known. Bai and Ng (2002) showed that the number of factors can be found by minimizing the following:

$$IC(k) = \log\big(V(k)\big) + k\left(\frac{n+T}{nT}\right)\log\left(\frac{nT}{n+T}\right).$$

(2) Once the estimates of w_{it}, $\widehat{w}_{it} = (\widehat{F}_t', \hat{u}_{it}, \Delta x_{it}')'$, were estimated, we used

$$\widehat{\Sigma} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{w}_{it}\widehat{w}_{it}' \tag{1.10}$$

to estimate Σ, where

$$\hat{u}_{it} = \hat{e}_{it} - \hat{\lambda}_i'\widehat{F}_t.$$

Ω was estimated by

$$\begin{aligned}\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{N}\Bigg\{&\frac{1}{T}\sum_{t=1}^{T}\widehat{w}_{it}\widehat{w}_{it}' \\ &+ \frac{1}{T}\sum_{\tau=1}^{l}\varpi_{\tau l}\sum_{t=\tau+1}^{T}(\widehat{w}_{it}\widehat{w}_{it-\tau}' + \widehat{w}_{it-\tau}\widehat{w}_{it}')\Bigg\},\end{aligned} \tag{1.11}$$

where $\varpi_{\tau l}$ is a weight function or a kernel. Using Phillips and Moon (1999), $\widehat{\Sigma}_i$ and $\widehat{\Omega}_i$ can be shown to be consistent for Σ_i and Ω_i.

1.7 Hypothesis testing

We now consider a linear hypothesis that involves the elements of the coefficient vector β. We show that hypothesis tests constructed using the FM estimator have asymptotic chi-squared distributions. The null hypothesis has the form:

$$H_0\colon\ R\beta = r, \tag{1.12}$$

where r is a $m \times 1$ known vector and R is a known $m \times k$ matrix describing the restrictions. A natural test statistic of the Wald test using $\hat{\beta}_{\mathrm{FM}}$ is

$$W = \frac{1}{6} n T^2 (R\hat{\beta}_{\mathrm{FM}} - r)' \Bigg[6 \widehat{\Omega}_{\varepsilon}^{-1} \Bigg\{ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\hat{\lambda}_i' \widehat{\Omega}_{F.\varepsilon i} \hat{\lambda}_i \widehat{\Omega}_{\varepsilon i} + \widehat{\Omega}_{u.\varepsilon i} \widehat{\Omega}_{\varepsilon i}) \Bigg\} \widehat{\Omega}_{\varepsilon}^{-1} \Bigg]^{-1} (R\hat{\beta}_{\mathrm{FM}} - r). \tag{1.13}$$

It is clear that W converges in distribution to a chi-squared random variable with k degrees of freedom, χ_k^2, as $(n, T \to \infty)$ under the null hypothesis. Hence, we establish the following theorem:

THEOREM 1.4. *If Assumptions 1.1–1.4 hold, then under the null hypothesis (1.12), with $(n, T \to \infty)$, $W \Rightarrow \chi_k^2$,*

REMARK 1.6. (1) One common application of Theorem 1.4 is the single-coefficient test: one of the coefficient is zero; $\beta_j = \beta_0$,

$$R = [0 \quad 0 \quad \cdots \quad 1 \quad 0 \quad \cdots \quad 0]$$

and $r = 0$. We can construct a t-statistic

$$t_j = \frac{\sqrt{n} T (\hat{\beta}_{j\mathrm{FM}} - \beta_0)}{s_j}, \tag{1.14}$$

where

$$s_j^2 = \Bigg[6 \widehat{\Omega}_{\varepsilon}^{-1} \Bigg\{ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\hat{\lambda}_i' \widehat{\Omega}_{F.\varepsilon i} \hat{\lambda}_i \widehat{\Omega}_{\varepsilon i} + \widehat{\Omega}_{u.\varepsilon i} \widehat{\Omega}_{\varepsilon i}) \Bigg\} \widehat{\Omega}_{\varepsilon}^{-1} \Bigg]_{jj},$$

the jth diagonal element of

$$\Bigg[6 \widehat{\Omega}_{\varepsilon}^{-1} \Bigg\{ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\hat{\lambda}_i' \widehat{\Omega}_{F.\varepsilon i} \hat{\lambda}_i \widehat{\Omega}_{\varepsilon i} + \widehat{\Omega}_{u.\varepsilon i} \widehat{\Omega}_{\varepsilon i}) \Bigg\} \widehat{\Omega}_{\varepsilon}^{-1} \Bigg].$$

It follows that

$$t_j \Rightarrow N(0, 1). \tag{1.15}$$

(2) General nonlinear parameter restriction such as H_0: $h(\beta) = 0$, where $h(\cdot)$, is $k^* \times 1$ vector of smooth functions such that $\frac{\partial h}{\partial \beta'}$ has full rank k^* can be conducted in a similar fashion as in Theorem 1.4. Thus, the Wald test has the following form

$$W_h = nT^2 h(\hat{\beta}_{\mathrm{FM}})' \widehat{V}_h^{-1} h(\hat{\beta}_{\mathrm{FM}}),$$

where

$$\widehat{V}_h^{-1} = \left(\frac{\partial h(\hat{\beta}_{\mathrm{FM}})}{\partial \beta'}\right) \widehat{V}_\beta^{-1} \left(\frac{\partial h(\hat{\beta}'_{\mathrm{FM}})}{\partial \beta}\right)$$

and

$$\widehat{V}_\beta = 6\widehat{\Omega}_\varepsilon^{-1} \left\{ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\hat{\lambda}_i' \widehat{\Omega}_{F.\varepsilon i} \hat{\lambda} \widehat{\Omega}_{\varepsilon ii} + \widehat{\Omega}_{u.\varepsilon i} \widehat{\Omega}_{\varepsilon i}) \right\} \widehat{\Omega}_\varepsilon^{-1}. \tag{1.16}$$

It follows that

$$W_h \Rightarrow \chi^2_{k^*}$$

as $(n, T \to \infty)$.

1.8 Monte Carlo simulations

In this section, we conduct Monte Carlo experiments to assess the finite sample properties of OLS and FM estimators. The simulations were performed by a Sun SparcServer 1000 and an Ultra Enterprise 3000. GAUSS 3.2.31 and COINT 2.0 were used to perform the simulations. Random numbers for error terms, $(F_t^*, u_{it}^*, \varepsilon_{it}^*)$ were generated by the GAUSS procedure RNDNS. At each replication, we generated an $n(T + 1000)$ length of random numbers and then split it into n series so that each series had the same mean and variance. The first 1,000 observations were discarded for each series. $\{F_t^*\}$, $\{u_{it}^*\}$ and $\{\varepsilon_{it}^*\}$ were constructed with $F_t^* = 0$, $u_{i0}^* = 0$ and $\varepsilon_{i0}^* = 0$.

To compare the performance of the OLS and FM estimators we conducted Monte Carlo experiments based on a design which is similar to Kao and Chiang (2000)

$$y_{it} = \alpha_i + \beta x_{it} + e_{it},$$

$$e_{it} = \lambda_i' F_t + u_{it},$$

and

$$x_{it} = x_{it-1} + \varepsilon_{it}$$

for $i = 1, \ldots, n$, $t = 1, \ldots, T$, where

$$\begin{pmatrix} F_t \\ u_{it} \\ \varepsilon_{it} \end{pmatrix} = \begin{pmatrix} F_t^* \\ u_{it}^* \\ \varepsilon_{it}^* \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.3 & -0.4 \\ \theta_{31} & \theta_{32} & 0.6 \end{pmatrix} \begin{pmatrix} F_{t-1}^* \\ u_{it-1}^* \\ \varepsilon_{it-1}^* \end{pmatrix} \tag{1.17}$$

with

$$\begin{pmatrix} F_t^* \\ u_{it}^* \\ \varepsilon_{it}^* \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & 1 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & 1 \end{bmatrix} \right).$$

For this experiment, we have a single factor ($r = 1$) and λ_i are generated from i.i.d. $N(\mu_\lambda, 1)$. We let $\mu_\lambda = 0.1$. We generated α_i from a uniform distribution, $U[0, 10]$, and set $\beta = 2$. From Theorems 1.1–1.3 we know that the asymptotic results depend upon variances and covariances of F_t, u_{it} and ε_{it}. Here we set $\sigma_{12} = 0$. The design in (1.17) is a good one since the endogeneity of the system is controlled by only four parameters, θ_{31}, θ_{32}, σ_{31} and σ_{32}. We choose $\theta_{31} = 0.8$, $\theta_{32} = 0.4$, $\sigma_{31} = -0.8$ and $\theta_{32} = 0.4$.

The estimate of the long-run covariance matrix in (1.11) was obtained by using the procedure KERNEL in COINT 2.0 with a Bartlett window. The lag truncation number was set arbitrarily at five. Results with other kernels, such as Parzen and quadratic spectral kernels, are not reported, because no essential differences were found for most cases.

Next, we recorded the results from our Monte Carlo experiments that examined the finite-sample properties of (a) the OLS estimator, $\hat{\beta}_{\text{OLS}}$ in (1.6), (b) the 2S-FM estimator, $\hat{\beta}_{2\text{S}}$, in (1.8), (c) the two-step naive FM estimator, $\hat{\beta}_{\text{FM}}^b$, proposed by Kao and Chiang (2000) and Phillips and Moon (1999), (d) the CUP-FM estimator $\hat{\beta}_{\text{CUP}}$, in (1.9) and (e) the CUP naive FM estimator $\hat{\beta}_{\text{FM}}^d$ which is similar to the two-step naive FM except the iteration goes beyond two steps. The naive FM estimators are obtained assuming the cross-sectional independence. The maximum number of the iteration for CUP-FM estimators is set to 20. The results we report are based on 1,000 replications and are summarized in Tables 1.1–1.4. All the FM estimators were obtained by using a Bartlett window of lag length five as in (1.11).

Table 1.1 reports the Monte Carlo means and standard deviations (in parentheses) of $(\hat{\beta}_{\text{OLS}} - \beta)$, $(\hat{\beta}_{2\text{S}} - \beta)$, $(\hat{\beta}_{\text{FM}}^b - \beta)$, $(\hat{\beta}_{\text{CUP}} - \beta)$, and $(\hat{\beta}_{\text{FM}}^d - \beta)$ for sample sizes $T = n = (20, 40, 60)$. The biases of the OLS estimator, $\hat{\beta}_{\text{OLS}}$, decrease at a rate of T. For example, with $\sigma_\lambda = 1$ and $\sigma_F = 1$, the bias at $T = 20$ is -0.045, at $T = 40$ is -0.024, and at $T = 60$ is -0.015. Also, the biases stay the same for different values of σ_λ and σ_F.

While we expected the OLS estimator to be biased, we expected FM estimators to produce better estimates. However, it is noticeable that the

Table 1.1. Means biases and standard deviation of OLS and FM estimators

	$\sigma_\lambda = 1$					$\sigma_\lambda = \sqrt{10}$					$\sigma_\lambda = \sqrt{0.5}$				
	OLS	FM[a]	FM[b]	FM[c]	FM[d]	OLS	FM[a]	FM[b]	FM[c]	FM[d]	OLS	FM[a]	FM[b]	FM[c]	FM[d]
$\sigma_F = 1$															
$T = 20$	−0.045	−0.025	−0.029	−0.001	−0.006	−0.046	−0.025	−0.029	−0.001	−0.006	−0.045	−0.025	−0.029	−0.001	−0.006
	(0.029)	(0.028)	(0.029)	(0.034)	(0.030)	(0.059)	(0.054)	(0.059)	(0.076)	(0.060)	(0.026)	(0.026)	(0.026)	(0.030)	(0.028)
$T = 40$	−0.024	−0.008	−0.011	−0.002	−0.005	−0.024	−0.009	−0.012	−0.003	−0.005	−0.024	−0.008	−0.011	−0.002	−0.005
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.020)	(0.019)	(0.019)	(0.021)	(0.018)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
$T = 60$	−0.015	−0.004	−0.005	−0.001	−0.003	−0.015	−0.003	−0.005	−0.001	−0.002	−0.015	−0.004	−0.005	−0.001	−0.003
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.011)	(0.010)	(0.010)	(0.011)	(0.010)	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)
$\sigma_F = \sqrt{10}$															
$T = 20$	−0.054	−0.022	−0.036	0.011	−0.005	−0.057	−0.024	−0.038	0.013	−0.003	−0.054	−0.022	−0.036	0.011	−0.005
	(0.061)	(0.054)	(0.061)	(0.078)	(0.062)	(0.176)	(0.156)	(0.177)	(0.228)	(0.177)	(0.046)	(0.042)	(0.047)	(0.059)	(0.047)
$T = 40$	−0.028	−0.007	−0.015	0.001	−0.007	−0.030	−0.009	−0.017	−0.001	−0.009	−0.028	−0.007	−0.014	0.001	−0.007
	(0.021)	(0.019)	(0.019)	(0.021)	(0.019)	(0.059)	(0.054)	(0.057)	(0.061)	(0.053)	(0.016)	(0.015)	(0.015)	(0.016)	(0.015)
$T = 60$	−0.018	−0.002	−0.007	0.001	−0.004	−0.017	−0.001	−0.006	0.002	−0.003	−0.018	−0.002	−0.007	0.001	−0.004
	(0.011)	(0.011)	(0.011)	(0.011)	(0.010)	(0.032)	(0.029)	(0.030)	(0.031)	(0.029)	(0.009)	(0.008)	(0.008)	(0.009)	(0.008)
$\sigma_F = \sqrt{0.5}$															
$T = 20$	−0.044	−0.025	−0.028	−0.003	−0.006	−0.045	−0.026	−0.028	−0.002	−0.006	−0.044	−0.026	−0.028	−0.003	−0.006
	(0.026)	(0.026)	(0.026)	(0.030)	(0.028)	(0.045)	(0.041)	(0.045)	(0.056)	(0.046)	(0.024)	(0.025)	(0.025)	(0.028)	(0.026)
$T = 40$	−0.023	−0.009	−0.010	−0.003	−0.004	−0.023	−0.009	−0.011	−0.003	−0.005	−0.023	−0.009	−0.010	−0.003	−0.004
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.016)	(0.015)	(0.015)	(0.016)	(0.014)	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)
$T = 60$	−0.015	−0.004	−0.005	−0.001	−0.003	−0.015	−0.004	−0.005	−0.001	−0.002	−0.015	−0.004	−0.005	−0.001	−0.003
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)

Note: (a) FM[a] is the 2S-FM, FM[b] is the naive 2S-FM, FM[c] is the CUP-FM and FM[d] is the naive CUP-FM. (b) $\mu_\lambda = 0.1$, $\sigma_{31} = -0.8$, $\sigma_{21} = -0.4$, $\theta_{31} = 0.8$, and $\theta_{21} = 0.4$.

Table 1.2. Means biases and standard deviation of OLS and FM estimators for different n and T

(n, T)	OLS	FM[a]	FM[b]	FM[c]	FM[d]
(20, 20)	−0.045	−0.019	−0.022	−0.001	−0.006
	(0.029)	(0.028)	(0.029)	(0.034)	(0.030)
(20, 40)	−0.024	−0.006	−0.009	−0.001	−0.004
	(0.014)	(0.014)	(0.013)	(0.014)	(0.013)
(20, 60)	−0.017	−0.004	−0.006	−0.001	−0.003
	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)
(20, 120)	−0.008	−0.001	−0.002	−0.000	−0.001
	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)
(40, 20)	−0.044	−0.018	−0.021	−0.002	−0.006
	(0.021)	(0.019)	(0.019)	(0.023)	(0.021)
(40, 40)	−0.024	−0.007	−0.009	−0.002	−0.004
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
(40, 60)	−0.015	−0.003	−0.005	−0.001	−0.002
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
(40, 120)	−0.008	−0.001	−0.002	−0.001	−0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
(60, 20)	−0.044	−0.018	−0.022	−0.002	−0.007
	(0.017)	(0.016)	(0.016)	(0.019)	(0.017)
(60, 40)	−0.022	−0.006	−0.008	−0.002	−0.004
	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)
(60, 60)	−0.015	−0.003	−0.005	−0.001	−0.003
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
(60, 120)	−0.008	−0.001	−0.002	−0.001	−0.001
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
(120, 20)	−0.044	−0.018	−0.022	−0.002	−0.007
	(0.013)	(0.011)	(0.012)	(0.013)	(0.012)
(120, 40)	−0.022	−0.006	−0.008	−0.002	−0.004
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
(120, 60)	−0.015	−0.003	−0.005	−0.001	−0.003
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
(120, 120)	−0.008	−0.001	−0.002	−0.001	−0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

Note: $\mu_\lambda = 0.1$, $\sigma_{31} = -0.8$, $\sigma_{21} = -0.4$, $\theta_{31} = 0.8$, and $\theta_{21} = 0.4$.

2S-FM estimator still has a downward bias for all values of σ_λ and σ_F, though the biases are smaller. In general, the 2S-FM estimator presents the same degree of difficulty with bias as does the OLS estimator. This is probably due to the failure of the nonparametric correction procedure.

In contrast, the results in Table 1.1 show that the CUP-FM, is distinctly superior to the OLS and 2S-FM estimators for all cases in terms of the mean biases. Clearly, the CUP-FM outperforms both the OLS and 2S-FM estimators.

Table 1.3. Means biases and standard deviation of t-statistics

	$\sigma_\lambda = 1$					$\sigma_\lambda = \sqrt{10}$					$\sigma_\lambda = \sqrt{0.5}$				
	OLS	FM^a	FM^b	FM^c	FM^d	OLS	FM^a	FM^b	FM^c	FM^d	OLS	FM^a	FM^b	FM^c	FM^d
$\sigma_F = 1$															
$T = 20$	−1.994	−1.155	−1.518	−0.056	−0.285	−0.929	−0.546	−0.813	−0.006	−0.122	−2.248	−1.299	−1.656	−0.071	−0.321
	(1.205)	(1.267)	(1.484)	(1.283)	(1.341)	(1.149)	(1.059)	(1.495)	(1.205)	(1.254)	(1.219)	(1.325)	(1.490)	(1.314)	(1.366)
$T = 40$	−2.915	−0.941	−1.363	−0.227	−0.559	−1.355	−0.465	−0.766	−0.128	−0.326	−3.288	−1.056	−1.474	−0.250	−0.602
	(1.202)	(1.101)	(1.248)	(1.054)	(1.141)	(1.127)	(0.913)	(1.207)	(0.912)	(1.049)	(1.221)	(1.151)	(1.253)	(1.096)	(1.159)
$T = 60$	−3.465	−0.709	−1.158	−0.195	−0.574	−1.552	−0.308	−0.568	−0.074	−0.261	−3.926	−0.814	−1.280	−0.229	−0.643
	(1.227)	(1.041)	(1.177)	(0.996)	(1.100)	(1.146)	(0.868)	(1.113)	(0.851)	(1.016)	(1.244)	(1.091)	(1.189)	(1.042)	(1.118)
$\sigma_F = \sqrt{10}$															
$T = 20$	−1.078	−0.484	−0.984	0.180	−0.096	−0.373	−0.154	−0.350	0.085	−0.006	−1.427	−0.639	1.257	0.229	−0.138
	(1.147)	(1.063)	(1.501)	(1.220)	(1.271)	(1.119)	(0.987)	(1.508)	(1.194)	(1.223)	(1.163)	(1.117)	(1.498)	(1.244)	(1.301)
$T = 40$	−1.575	−0.355	−0.963	0.042	−0.407	−0.561	−0.152	−0.397	−0.014	−0.190	−2.082	−0.453	−1.211	0.073	−0.506
	(1.131)	(0.917)	(1.214)	(0.926)	(1.063)	(1.097)	(0.844)	(1.179)	(0.871)	(1.008)	(1.154)	(0.967)	(1.232)	(0.967)	(1.096)
$T = 60$	−1.809	−0.155	−0.776	0.111	−0.390	−0.588	−0.041	−0.247	0.049	−0.111	−2.424	−0.212	−1.019	0.143	−0.523
	(1.158)	(0.879)	(1.131)	(0.867)	(1.035)	(1.108)	(0.812)	(1.078)	(0.811)	(0.983)	(1.192)	(0.929)	(1.162)	(0.909)	(1.069)
$\sigma_F = \sqrt{0.5}$															
$T = 20$	−2.196	−1.319	−1.606	−0.137	−0.327	−1.203	−0.734	−1.008	−0.054	−0.176	−2.367	−1.421	−1.692	−0.157	−0.351
	(1.219)	(1.325)	(1.488)	(1.307)	(1.362)	(1.164)	(1.112)	(1.488)	(1.217)	(1.273)	(1.231)	(1.363)	(1.492)	(1.324)	(1.379)
$T = 40$	−3.214	−1.093	−1.415	−0.311	−0.576	−1.752	−0.619	−0.922	−0.188	−0.385	−3.462	−1.176	−1.481	−0.333	−0.599
	(1.226)	(1.057)	(1.155)	(1.104)	(1.169)	(1.148)	(0.962)	(1.222)	(0.944)	(1.087)	(1.236)	(1.185)	(1.255)	(1.121)	(1.168)
$T = 60$	−3.839	−0.868	−1.217	−0.296	−0.602	−2.037	−0.446	−0.712	−0.139	−0.331	−4.149	−0.949	−1.295	−0.329	−0.646
	(1.239)	(1.088)	(1.183)	(1.037)	(1.112)	(1.169)	(0.908)	(1.131)	(0.881)	(1.038)	(1.249)	(1.123)	(1.190)	(1.069)	(1.122)

Note: (a) FM^a is the 2S-FM, FM^b is the naive 2S-FM, FM^c is the CUP-FM and FM^d is the naive CUP-FM. (b) $\mu_\lambda = 0.1$, $\sigma_{31} = -0.8$, $\sigma_{21} = -0.4$, $\theta_{31} = 0.8$, and $\theta_{21} = 0.4$.

Table 1.4. Means biases and standard deviation of t-statistics for different n and T

(n, T)	OLS	FM[a]	FM[b]	FM[c]	FM[d]
(20, 20)	−1.994	−0.738	−1.032	−0.056	−0.286
	(1.205)	(1.098)	(1.291)	(1.283)	(1.341)
(20, 40)	−2.051	−0.465	−0.725	−0.105	−0.332
	(1.179)	(0.999)	(1.126)	(1.046)	(1.114)
(20, 60)	−2.129	−0.404	−0.684	−0.162	−0.421
	(1.221)	(0.963)	(1.278)	(0.983)	(1.111)
(20, 120)	−2.001	−0.213	−0.456	−0.095	−0.327
	(1.222)	(0.923)	(1.083)	(0.931)	(1.072)
(40, 20)	−2.759	−1.017	−1.404	−0.103	−0.402
	(1.237)	(1.116)	(1.291)	(1.235)	(1.307)
(40, 40)	−2.915	−0.699	−1.075	−0.227	−0.559
	(1.202)	(1.004)	(1.145)	(1.054)	(1.141)
(40, 60)	−2.859	−0.486	−0.835	−0.173	−0.493
	(1.278)	(0.998)	(1.171)	(1.014)	(1.154)
(40, 120)	−2.829	−0.336	−0.642	−0.181	−0.472
	(1.209)	(0.892)	(1.047)	(0.899)	(1.037)
(60, 20)	−3.403	−1.252	−1.740	−0.152	−0.534
	(1.215)	(1.145)	(1.279)	(1.289)	(1.328)
(60, 40)	−3.496	−0.807	−1.238	−0.255	−0.635
	(1.247)	(1.016)	(1.165)	(1.053)	(1.155)
(60, 60)	−3.465	−0.573	−0.987	−0.195	−0.574
	(1.227)	(0.974)	(1.111)	(0.996)	(1.100)
(60, 120)	−3.515	−0.435	−0.819	−0.243	−0.609
	(1.197)	(0.908)	(1.031)	(0.913)	(1.020)
(120, 20)	−4.829	−1.758	−2.450	−0.221	−0.760
	(1.345)	(1.162)	(1.327)	(1.223)	(1.308)
(120, 40)	−4.862	−1.080	−1.679	−0.307	−0.831
	(1.254)	(1.022)	(1.159)	(1.059)	(1.143)
(120, 60)	−4.901	−0.852	−1.419	−0.329	−0.846
	(1.239)	(0.964)	(1.097)	(0.978)	(1.077)
(120, 120)	−5.016	−0.622	−1.203	−0.352	−0.908
	(1.248)	(0.922)	(1.059)	(0.927)	(1.048)

Note: $\mu_\lambda = 0.1$, $\sigma_{31} = -0.8$, $\sigma_{21} = -0.4$, $\theta_{31} = 0.8$, and $\theta_{21} = 0.4$.

It is important to know the effects of the variations in panel dimensions on the results, since the actual panel data have a wide variety of cross-section and time-series dimensions. Table 1.2 considers 16 different combinations for n and T, each ranging from 20 to 120 with $\sigma_{31} = -0.8$, $\sigma_{21} = -0.4$, $\theta_{31} = 0.8$, and $\theta_{21} = 0.4$. First, we notice that the cross-section dimension has no significant effect on the biases of all estimators. From this it seems that in practice the T dimension must exceed the n

dimension, especially for the OLS and 2S-FM estimators, in order to get a good approximation of the limiting distributions of the estimators. For example, for OLS estimator in Table 1.2, the reported bias, -0.008, is substantially less for $(T = 120, n = 40)$ than it is for either $(T = 40, n = 40)$ (the bias is -0.024), or $(T = 40, n = 120)$ (the bias is -0.022). The results in Table 1.2 again confirm the superiority of the CUP-FM.

Monte Carlo means and standard deviations of the t-statistic, $t_{\beta=\beta_0}$, are given in Table 1.3. Here, the OLS t-statistic is the conventional t-statistic as printed by standard statistical packages. With all values of σ_λ and σ_F with the exception $\sigma_\lambda = \sqrt{10}$, the CUP-FM t-statistic is well approximated by a standard $N(0, 1)$ suggested from the asymptotic results. The CUP-FM t-statistic is much closer to the standard normal density than the OLS t-statistic and the 2S-FM t-statistic. The 2S-FM t-statistic is not well approximated by a standard $N(0, 1)$.

Table 1.4 shows that both the OLS t-statistic and the FM t-statistics become more negatively biased as the dimension of cross-section n increases. The heavily negative biases of the 2S-FM t-statistic in Tables 1.3–1.4 again indicate the poor performance of the 2S-FM estimator. For the CUP-FM, the biases decrease rapidly and the standard errors converge to 1.0 as T increases.

It is known that when the length of time series is short the estimate $\widehat{\Omega}$ in (1.11) may be sensitive to the length of the bandwidth. In Tables 1.2 and 1.4, we first investigate the sensitivity of the FM estimators with respect to the choice of length of the bandwidth. We extend the experiments by changing the lag length from 5 to other values for a Barlett window. Overall, the results (not reported here) show that changing the lag length from 5 to other values does not lead to substantial changes in biases for the FM estimators and their t-statistics.

1.9 Conclusion

A factor approach to panel models with cross-sectional dependence is useful when both the time series and cross-sectional dimensions are large. This approach also provides significant reduction in the number of variables that may cause the cross-sectional dependence in panel data. In this paper, we study the estimation and inference of a panel cointegration model with cross-sectional dependence. The paper contributes to the growing literature on panel data with cross-sectional dependence by (i) discussing limiting distributions for the OLS and FM estimators, (ii) suggesting a CUP-FM estimator and (iii) investigating the finite sample proprieties of the OLS, CUP-FM and 2S-FM estimators. It is found that the 2S-FM and OLS estimators have a nonnegligible bias in finite samples, and that the CUP-FM estimator improves over the other two estimators.

Acknowledgements

We thank Badi Baltagi, Yu-Pin Hu, Giovanni Urga, Kamhon Kan, Chung-Ming Kuan, Hashem Pesaran, Lorenzo Trapani and Yongcheol Shin for helpful comments. We also thank seminar participants at Academia Sinica, National Taiwan University, Syracuse University, Workshop on Recent Developments in the Econometrics of Panel Data in London, March 2004 and the European Meeting of the Econometric Society in Madrid, August 2004 for helpful comments and suggestions. Jushan Bai acknowledges financial support from the NSF (grant SES-0137084).

Appendix A1

Let

$$B_{nT}=\left[\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\right].$$

Note

$$\begin{aligned}\sqrt{n}T(\hat{\beta}_{\text{OLS}}-\beta)&\\ &=\left[\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}e_{it}(x_{it}-\bar{x}_i)'\right)\right]\left[\frac{1}{n}\frac{1}{T^2}B_{nT}\right]^{-1}\\ &=\left[\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\zeta_{1iT}\right]\left[\frac{1}{n}\sum_{i=1}^{n}\zeta_{2iT}\right]^{-1}\\ &=\sqrt{n}\xi_{1nT}[\xi_{2nT}]^{-1},\end{aligned}$$

where $\bar{x}_i=\frac{1}{T}\sum_{t=1}^{T}x_{it}$, $\bar{y}_i=\frac{1}{T}\sum_{t=1}^{T}y_{it}$, $\zeta_{1iT}=\frac{1}{T}\sum_{t=1}^{T}e_{it}(x_{it}-\bar{x}_i)'$, $\zeta_{2iT}=\frac{1}{T^2}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'$, $\xi_{1nT}=\frac{1}{n}\sum_{i=1}^{n}\zeta_{1iT}$, and $\xi_{2nT}=\frac{1}{n}\sum_{i=1}^{n}\zeta_{2iT}$. Before going into the next theorem, we need to consider some preliminary results.

Define $\Omega_\varepsilon=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Omega_{\varepsilon i}$ and

$$\begin{aligned}\theta^n=\frac{1}{n}\Bigg[\sum_{i=1}^{n}\lambda_i'\Bigg(\Omega_{F.\varepsilon i}\Omega_{\varepsilon i}^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}x_{it}'(x_{it}-\bar{x}_i)\right)\Omega_{\varepsilon i}^{1/2}+\Delta_{F\varepsilon i}\Bigg)\\ +\Omega_{u.\varepsilon i}\Omega_{\varepsilon i}^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}x_{it}'(x_{it}-\bar{x}_i)\right)\Omega_{\varepsilon i}^{1/2}+\Delta_{u\varepsilon i}\Bigg].\end{aligned}$$

If Assumptions 1.1–1.4 hold, then

LEMMA A1.1. (a) *As* $(n, T \to \infty)$,

$$\frac{1}{n}\frac{1}{T^2}B_{nT} \xrightarrow{p} \frac{1}{6}\Omega_{\varepsilon}.$$

(b) *As* $(n, T \to \infty)$ *with* $\frac{n}{T} \to 0$,

$$\begin{aligned}
&\sqrt{n}\left(\frac{1}{n}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T} e_{it}(x_{it}-\bar{x}_i)' - \theta^n\right)\\
&\quad \Rightarrow N\left(0, \frac{1}{6}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\{\lambda_i'\Omega_{F.\varepsilon i}\lambda_i\Omega_{\varepsilon i} + \Omega_{u.\varepsilon i}\Omega_{\varepsilon i}\}\right).
\end{aligned}$$

PROOF. (a) and (b) can be shown easily by following Theorem 8 in Phillips and Moon (1999). □

A1.1 Proof of Theorem 1.1

PROOF. Recall that

$$\begin{aligned}
&\sqrt{n}T(\hat{\beta}_{\mathrm{OLS}} - \beta) - \sqrt{n}\frac{1}{n}\left[\sum_{i=1}^{n}\lambda_i'\left(\Omega_{F\varepsilon i}\Omega_{\varepsilon i}^{-1/2}\right.\right.\\
&\qquad \times \left(\frac{1}{T}\sum_{t=1}^{T}x_{it}'(x_{it}-\bar{x}_i)\right)\Omega_{\varepsilon i}^{1/2} + \Delta_{F\varepsilon i}\Bigg)\\
&\qquad \left. + \Omega_{\varepsilon ui}\Omega_{\varepsilon i}^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}x_{it}'(x_{it}-\bar{x}_i)\right)\Omega_{\varepsilon i}^{1/2} + \Delta_{\varepsilon ui}\right]\left[\frac{1}{n}\frac{1}{T^2}B_{nT}\right]^{-1}\\
&\quad = \left[\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left\{\zeta_{1iT} - \lambda_i'\left(\Omega_{F\varepsilon i}\Omega_{\varepsilon i}^{-1/2}\right.\right.\right.\\
&\qquad \times \left(\frac{1}{T}\sum_{t=1}^{T}x_{it}'(x_{it}-\bar{x}_i)\right)\Omega_{\varepsilon i}^{1/2} + \Delta_{F\varepsilon i}\Bigg)\\
&\qquad \left.\left. - \Omega_{u\varepsilon i}\Omega_{\varepsilon i}^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}x_{it}'(x_{it}-\bar{x}_i)\right)\Omega_{\varepsilon i}^{1/2} + \Delta_{u\varepsilon i}\right\}\right]\\
&\qquad \times \left[\frac{1}{n}\sum_{i=1}^{n}\zeta_{2iT}\right]^{-1}\\
&\quad = \left[\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\zeta_{1iT}^{*}\right]\left[\frac{1}{n}\sum_{i=1}^{n}\zeta_{2iT}\right]^{-1}\\
&\quad = \sqrt{n}\xi_{1nT}^{*}[\xi_{2nT}]^{-1},
\end{aligned}$$

where

$$\zeta^*_{1iT} = \zeta_{1iT} - \lambda'_i\left(\varOmega_{F\varepsilon i}\varOmega_{\varepsilon i}^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}x'_{it}(x_{it}-\bar{x}_i)\right)\varOmega_{\varepsilon i}^{1/2}+\varDelta_{F\varepsilon i}\right)$$
$$-\varOmega_{u\varepsilon i}\varOmega_{\varepsilon i}^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}x'_{it}(x_{it}-\bar{x}_i)\right)\varOmega_{\varepsilon i}^{1/2}+\varDelta_{u\varepsilon i}$$

and

$$\xi^*_{1nT} = \frac{1}{n}\sum_{i=1}^{n}\zeta^*_{1iT}.$$

First, we note from Lemma A1.1(b) that

$$\sqrt{n}\xi^*_{1nT} \Rightarrow N\left(0, \frac{1}{6}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\{\lambda'_i\varOmega_{F.\varepsilon i}\lambda_i\varOmega_{\varepsilon i}+\varOmega_{u.\varepsilon i}\varOmega_{\varepsilon i}\}\right)$$

as $(n, T \to \infty)$ and $\frac{n}{T} \to 0$. Using the Slutsky theorem and (a) from Lemma A1.1, we obtain

$$\sqrt{n}\xi^*_{1nT}[\xi_{2nT}]^{-1}$$
$$\Rightarrow N\left(0, 6\varOmega_{\varepsilon}^{-1}\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\lambda'_i\varOmega_{F.\varepsilon i}\lambda_i\varOmega_{\varepsilon i}+\varOmega_{u.\varepsilon i}\varOmega_{\varepsilon i})\right\}\varOmega_{\varepsilon}^{-1}\right).$$

Hence,

$$\sqrt{n}T(\hat{\beta}_{\text{OLS}}-\beta)-\sqrt{n}\delta_{nT}$$
$$\Rightarrow N\left(0, 6\varOmega_{\varepsilon}^{-1}\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\lambda'_i\varOmega_{F.\varepsilon i}\lambda_i\varOmega_{\varepsilon i}+\varOmega_{u.\varepsilon i}\varOmega_{\varepsilon i})\right\}\varOmega_{\varepsilon}^{-1}\right), \tag{A1.1}$$

proving the theorem, where

$$\delta_{nT} = \frac{1}{n}\left[\sum_{i=1}^{n}\lambda'_i\left(\varOmega_{F\varepsilon i}\varOmega_{\varepsilon i}^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}x'_{it}(x_{it}-\bar{x}_i)\right)\varOmega_{\varepsilon i}^{1/2}+\varDelta_{F\varepsilon i}\right)\right.$$
$$\left.+\varOmega_{u\varepsilon i}\varOmega_{\varepsilon i}^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}x'_{it}(x_{it}-\bar{x}_i)\right)\varOmega_{\varepsilon i}^{1/2}+\varDelta_{u\varepsilon i}\right]$$
$$\times\left[\frac{1}{n}\frac{1}{T^2}B_{nT}\right]^{-1}.$$

Therefore, we established Theorem 1.1. □

A1.2 Proof of Theorem 1.2

PROOF. Let

$$F_{it}^{+} = F_t - \Omega_{F\varepsilon i}\Omega_{\varepsilon i}^{-1}\varepsilon_{it},$$

and

$$u_{it}^{+} = u_{it} - \Omega_{u\varepsilon i}\Omega_{\varepsilon i}^{-1}\varepsilon_{it}.$$

The FM estimator of β can be rewritten as follows

$$\begin{aligned}\tilde{\beta}_{\mathrm{FM}} &= \left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T} y_{it}^{+}(x_{it}-\bar{x}_i)' - T\left(\lambda_i'\Delta_{F\varepsilon i}^{+} + \Delta_{u\varepsilon i}^{+}\right)\right)\right]B_{nT}^{-1}\\ &= \beta + \left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\left(\lambda_i'F_{it}^{+} + u_{it}^{+}\right)(x_{it}-\bar{x}_i)'\right.\right.\\ &\qquad \left.\left. - T\left(\lambda_i'\Delta_{F\varepsilon i}^{+} + \Delta_{u\varepsilon i}^{+}\right)\right)\right]B_{nT}^{-1}. \end{aligned} \tag{A1.2}$$

First, we rescale $(\tilde{\beta}_{\mathrm{FM}} - \beta)$ by $\sqrt{n}T$

$$\begin{aligned}\sqrt{n}T(\tilde{\beta}_{\mathrm{FM}} - \beta) &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\left[\left(\lambda_i'F_{it}^{+} + u_{it}^{+}\right)(x_{it}-\bar{x}_i)'\right.\\ &\qquad \left. - \lambda_i'\Delta_{F\varepsilon i}^{+} - \Delta_{u\varepsilon i}^{+}\right]\left[\frac{1}{n}\frac{1}{T^2}B_{nT}\right]^{-1}\\ &= \left[\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\zeta_{1iT}^{**}\right]\left[\frac{1}{n}\sum_{i=1}^{n}\zeta_{2iT}\right]^{-1}\\ &= \sqrt{n}\xi_{1nT}^{**}[\xi_{2nT}]^{-1}, \end{aligned} \tag{A1.3}$$

where $\zeta_{1iT}^{**} = \frac{1}{T}\sum_{t=1}^{T}[(\lambda_i'F_{it}^{+} + \hat{u}_{it}^{+})(x_{it}-\bar{x}_i)' - \lambda_i'\Delta_{F\varepsilon i}^{+} - \Delta_{u\varepsilon i}^{+}]$, and $\xi_{1nT}^{**} = \frac{1}{n}\sum_{i=1}^{n}\zeta_{1iT}^{**}$.

Modifying Theorem 11 in Phillips and Moon (1999) and Kao and Chiang (2000) we can show that as $(n, T \to \infty)$ with $\frac{n}{T} \to 0$

$$\begin{aligned}&\sqrt{n}\left(\frac{1}{n}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\lambda_i'F_{it}^{+}(x_{it}-\bar{x}_i)' - \lambda_i'\Delta_{F\varepsilon i}^{+}\right)\right)\\ &\quad \Rightarrow N\left(0, \frac{1}{6}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\lambda_i'\Omega_{F.\varepsilon i}\lambda_i\Omega_{\varepsilon i}\right)\end{aligned}$$

and

$$\sqrt{n}\left(\frac{1}{n}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\hat{u}_{it}^{+}(x_{it}-\bar{x}_i)'-\Delta_{u\varepsilon i}^{+}\right)\right)$$
$$\Rightarrow N\left(0,\frac{1}{6}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Omega_{u.\varepsilon i}\Omega_{\varepsilon i}\right)$$

and combing this with Assumption 1.4 that F_t and u_{it} are independent and Lemma A1.1(a) yields

$$\sqrt{n}T(\tilde{\beta}_{\mathrm{FM}}-\beta)$$
$$\Rightarrow N\left(0,6\Omega_{\varepsilon}^{-1}\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\lambda_i'\Omega_{F.\varepsilon i}\lambda_i\Omega_{\varepsilon i}+\Omega_{u.\varepsilon i}\Omega_{\varepsilon i})\right\}\Omega_{\varepsilon}^{-1}\right)$$

as required. □

A1.3 Proof of Lemma 1.3

PROOF. We note that λ_i is estimating $H\lambda_i$, and $\widehat{\Omega}_{F\varepsilon}$ is estimating $H^{-1'}\widehat{\Omega}_{F\varepsilon}$. Thus $\hat{\lambda}_i'\widehat{\Omega}_{F\varepsilon}$ is estimating $\lambda_i'\Omega_{F\varepsilon}$, which is the object of interest. For the purpose of notational simplicity, we shall assume H being a $r\times r$ identify matrix in our proof below. From

$$\hat{e}_{it}^{+}=e_{it}-(\hat{\lambda}_i'\widehat{\Omega}_{F\varepsilon}+\widehat{\Omega}_{u\varepsilon})\widehat{\Omega}_{\varepsilon}^{-1}\Delta x_{it}$$

and

$$e_{it}^{+}=e_{it}-(\lambda_i'\Omega_{F\varepsilon}+\Omega_{u\varepsilon})\Omega_{\varepsilon}^{-1}\Delta x_{it},$$

$$\begin{aligned}&\hat{e}_{it}^{+}-e_{it}^{+}\\&\quad=-\left[\left\{(\hat{\lambda}_i'\widehat{\Omega}_{F\varepsilon}+\widehat{\Omega}_{u\varepsilon})\widehat{\Omega}_{\varepsilon}^{-1}-(\lambda_i'\Omega_{F\varepsilon}+\Omega_{u\varepsilon})\Omega_{\varepsilon}^{-1}\right\}\Delta x_{it}\right]\\&\quad=-\left[\left\{\hat{\lambda}_i'\widehat{\Omega}_{F\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}-\lambda_i'\Omega_{F\varepsilon}\Omega_{\varepsilon}^{-1}+\widehat{\Omega}_{u\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}-\Omega_{u\varepsilon}\Omega_{\varepsilon}^{-1}\right\}\Delta x_{it}\right].\end{aligned}$$

Then,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\widehat{\Omega}_{u\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}-\Omega_{u\varepsilon}\Omega_{\varepsilon}^{-1}\right)\Delta x_{it}(x_{it}-\bar{x}_i)'\\&\quad=\left(\widehat{\Omega}_{u\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}-\Omega_{u\varepsilon}\Omega_{\varepsilon}^{-1}\right)\frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\Delta x_{it}(x_{it}-\bar{x}_i)'\\&\quad=\mathrm{o}_p(1)\mathrm{O}_p(1)\\&\quad=\mathrm{o}_p(1)\end{aligned}$$

because

$$\widehat{\Omega}_{u\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1} - \Omega_{u\varepsilon}\Omega_{\varepsilon}^{-1} = \mathrm{o}_p(1)$$

and

$$\frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\Delta x_{it}(x_{it} - \bar{x}_i)' = \mathrm{O}_p(1).$$

Thus

$$\begin{aligned}
&\frac{1}{\sqrt{n}T}\sum_{i=1}^{n}\sum_{t=1}^{T}\big(\hat{e}_{it}^{+} - e_{it}^{+}\big)(x_{it} - \bar{x}_i)' \\
&\quad = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\big[\big\{(\lambda_i'\Omega_{F\varepsilon} + \Omega_{u\varepsilon})\Omega_{\varepsilon}^{-1} \\
&\qquad - (\hat{\lambda}_i'\widehat{\Omega}_{F\varepsilon} + \widehat{\Omega}_{u\varepsilon})\widehat{\Omega}_{\varepsilon}^{-1}\big\}\Delta x_{it}\big](x_{it} - \bar{x}_i)' \\
&\quad = \frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\big(\lambda_i'\Omega_{F\varepsilon}\Omega_{\varepsilon}^{-1} - \hat{\lambda}_i'\widehat{\Omega}_{F\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}\big)\Delta x_{it}(x_{it} - \bar{x}_i)' \\
&\qquad + \frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\big(\Omega_{u\varepsilon}\Omega_{\varepsilon}^{-1} - \widehat{\Omega}_{u\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}\big)\Delta x_{it}(x_{it} - \bar{x}_i)' \\
&\quad = \frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\big(\lambda_i'\Omega_{F\varepsilon}\Omega_{\varepsilon}^{-1} - \hat{\lambda}_i'\widehat{\Omega}_{F\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}\big)\Delta x_{it}(x_{it} - \bar{x}_i)' \\
&\qquad + \mathrm{o}_p(1).
\end{aligned}$$

The remainder of the proof needs to show that

$$\frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\big(\lambda_i'\Omega_{F\varepsilon}\Omega_{\varepsilon}^{-1} - \hat{\lambda}_i'\widehat{\Omega}_{F\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}\big)\Delta x_{it}(x_{it} - \bar{x}_i)' = \mathrm{o}_p(1).$$

We write A for $\Omega_{F\varepsilon}\Omega_{\varepsilon}^{-1}$ and $\widehat{A}$ for $\widehat{\Omega}_{F\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}$ respectively and then

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\big(\lambda_i'\Omega_{F\varepsilon}\Omega_{\varepsilon}^{-1} - \hat{\lambda}_i'\widehat{\Omega}_{F\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}\big)\Delta x_{it}(x_{it} - \bar{x}_i)' \\
&\quad = \frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}(\lambda_i'A - \hat{\lambda}_i'\widehat{A})\Delta x_{it}(x_{it} - \bar{x}_i)' \\
&\quad = \frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\big[\lambda_i'(A - \widehat{A}) + (\lambda_i' - \hat{\lambda}_i')\widehat{A}\big]\Delta x_{it}(x_{it} - \bar{x}_i)'
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\lambda_i'(A-\widehat{A})\Delta x_{it}(x_{it}-\bar{x}_i)' \\
&\quad + \frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}(\lambda_i'-\hat{\lambda}_i')\widehat{A}\Delta x_{it}(x_{it}-\bar{x}_i)' \\
&= I + II, \quad \text{say.}
\end{aligned}
$$

Term I is a row vector. Let I_j be the jth component of I. Let ℓ_j be the jth column of an identity matrix so that $\ell_j = (0,\ldots,0,1,0,\ldots 0)'$. Left multiply I by ℓ_j to obtain the jth component, which is scalar and thus is equal to its trace. That is

$$
\begin{aligned}
I_j &= \mathrm{tr}\left[(A-\widehat{A})\left(\frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\lambda_i'\Delta x_{it}(x_{it}-\bar{x}_i)'\ell_j\lambda_i\right)\right] \\
&= \mathrm{tr}\big[(A-\widehat{A})O_p(1)\big] \\
&= \mathrm{o}_p(1)\mathrm{O}_p(1) \\
&= \mathrm{o}_p(1)
\end{aligned}
$$

because $\frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\lambda_i'\Delta x_{it}(x_{it}-\bar{x}_i)'\ell_j\lambda_i = \mathrm{O}_p(1)$ and $A-\widehat{A} = \mathrm{o}_p(1)$.

Next consider II. Let $c_i = \widehat{A}\frac{1}{T}\sum_{i=1}^{n}\Delta x_{it}(x_{it}-\bar{x}_i)'$. Then $c_i = \mathrm{O}_p(1)$ and $\frac{1}{n}\sum_{i=1}^{n}\|c_i\|^2 = \mathrm{O}_p(1)$, thus by Lemma 1.2(ii), we have

$$
\begin{aligned}
II &\leqslant \sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n}(\lambda_i'-\hat{\lambda}_i')c_i\right| \\
&= \sqrt{n}\mathrm{O}_p\left(\frac{1}{\delta_{nT}^2}\right) = \sqrt{n}\mathrm{O}_p\left(\frac{1}{\min[n,T]}\right) \\
&= \mathrm{O}_p\left(\frac{\sqrt{n}}{\min\{n,T\}}\right) \\
&= \mathrm{O}_p\left(\frac{1}{\sqrt{n}}\right) + \mathrm{O}_p\left(\frac{\sqrt{n}}{T}\right) \\
&= \mathrm{o}_p(1)
\end{aligned}
$$

since $(n,T\to\infty)$ and $\frac{\sqrt{n}}{T}\to 0$. This establishes

$$
\frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\big(\lambda_i'\Omega_{F\varepsilon}\Omega_{\varepsilon}^{-1}-\hat{\lambda}_i'\widehat{\Omega}_{F\varepsilon}\widehat{\Omega}_{\varepsilon}^{-1}\big)\Delta x_{it}(x_{it}-\bar{x}_i)' = \mathrm{o}_p(1)
$$

and proves Lemma 1.3. □

References

Andrews, D.W.K. (2005), "Cross-section regression with common shocks", *Econometrica*, Vol. 73, pp. 1551–1586.

Bai, J. (2004), "Estimating cross-section common stochastic trends in nonstationary panel data", *Journal of Econometrics*, Vol. 122, pp. 137–183.

Bai, J., Ng, S. (2002), "Determining the number of factors in approximate factor models", *Econometrica*, Vol. 70, pp. 191–221.

Bai, J., Ng, S. (2004), "A panic attack on unit roots and cointegration", *Econometrica*, Vol. 72, pp. 1127–1177.

Baltagi, B., Kao, C. (2000), "Nonstationary panels, cointegration in panels and dynamic panels: a survey", *Advances in Econometrics*, Vol. 15, pp. 7–51.

Baltagi, B., Song, S.H., Koh, W. (2004), "Testing panel data regression models with spatial error correlation", *Journal of Econometrics*, Vol. 117, pp. 123–150.

Chang, Y. (2002), "Nonlinear IV unit root tests in panels with cross-sectional dependency", *Journal of Econometrics*, Vol. 116, pp. 261–292.

Coakley, J., Fuerts, A., Smith, R.P. (2002), "A principal components approach to cross-section dependence in panels", Manuscript, Birckbeck College, University of London.

Forni, M., Reichlin, L. (1998), "Let's get real: a factor-analytic approach to disaggregated business cycle dynamics", *Review of Economic Studies*, Vol. 65, pp. 453–473.

Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2000), "Reference cycles: the NBER methodology revisited", CEPR Discussion Paper 2400.

Gregory, A., Head, A. (1999), "Common and country-specific fluctuations in productivity, investment, and the current account", *Journal of Monetary Economics*, Vol. 44, pp. 423–452.

Hall, S.G., Lazarova, S., Urga, G. (1999), "A principle component analysis of common stochastic trends in heterogeneous panel data: some Monte Carlo evidence", *Oxford Bulletin of Economics and Statistics*, Vol. 61, pp. 749–767.

Kao, C., Chiang, M.H. (2000), "On the estimation and inference of a cointegrated regression in panel data", *Advances in Econometrics*, Vol. 15, pp. 179–222.

Moon, H.R., Perron, B. (2004), "Testing for a unit root in panels with dynamic factors", *Journal of Econometrics*, Vol. 122, pp. 81–126.

Pesaran, H. (2004), "Estimation and inference in large heterogenous panels with a multifactor error structure", Manuscript, Trinity College, Cambridge.

Phillips, P.C.B., Hansen, B.E. (1990), "Statistical inference in instrumental variables regression with I(1) processes", *Review of Economic Studies*, Vol. 57, pp. 99–125.

Phillips, P.C.B., Moon, H. (1999), "Linear regression limit theory for nonstationary panel data", *Econometrica*, Vol. 67, pp. 1057–1111.

Phillips, P.C.B., Sul, D. (2003), "Dynamic panel estimation and homogeneity testing under cross section dependence", *Econometric Journal*, Vol. 6, pp. 217–259.

Robertson, D., Symon, J. (2000), "Factor residuals in SUR regressions: estimating panels allowing for cross sectional correlation", Manuscript, Faculty of Economics and Politics, University of Cambridge.

Stock, J.H., Watson, M.W. (2002), "Forecasting using principal components from a large number of predictors", *Journal of the American Statistical Association*, Vol. 97, pp. 1167–1179.

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74002-0

CHAPTER 2

A Full Heteroscedastic One-Way Error Components Model: Pseudo-Maximum Likelihood Estimation and Specification Testing

Bernard Lejeune

HEC-University of Liège, CORE and ERUDITE, Boulevard du Rectorat, 3, B33 4000 Liège, Belgium
E-mail address: B.Lejeune@ulg.ac.be

Abstract

This paper proposes an extension of the standard one-way error components model allowing for heteroscedasticity in both the individual-specific and the general error terms, as well as for unbalanced panel. On the grounds of its computational convenience, its potential efficiency, its robustness to non-normality and its robustness to possible misspecification of the assumed scedastic structure of the data, we argue for estimating this model by Gaussian pseudo-maximum likelihood of order two. Further, we review how, taking advantage of the powerful m-testing framework, the correct specification of the prominent aspects of the model may be tested. We survey potentially useful nested, non-nested, Hausman and information matrix type diagnostic tests of both the mean and the variance specification of the model. Finally, we illustrate the usefulness of our proposed model and estimation and diagnostic testing procedures through an empirical example.

Keywords: error components model, heteroscedasticity, unbalanced panel data, pseudo-maximum likelihood estimation, m-testing

JEL classifications: C12, C22, C52

2.1 Introduction

As largely acknowledged, heteroscedasticity is endemic when working with microeconomic cross-section data. One of its common sources is differences in size (the level of the variables) across individuals. This kind of heteroscedasticity is mechanical. It is simply a consequence of the additive disturbance structure of the classical regression model. It is generally tackled by performing a logarithmic transformation of the dependent variable. However, even after accounting in such a way for differences in size, numerous cases remain where we cannot expect the error variance to be constant. On one hand, there is a priori no reason to believe that the logarithmic specification postulating similar percentage variations across observations is relevant. In the production field for example, observations for lower outputs firms seem likely to evoke larger variances (see Baltagi and Griffin, 1988). On the other hand, the error variance may also vary across observations of similar size. For example, the variance of firms outputs might depend upon their capital intensity.

Obviously, there is no reason to expect the heteroscedasticity problems associated with microeconomic panel data to be markedly different from those encountered in work with cross-section data. Nonetheless, the issue of heteroscedasticity has received somewhat limited attention in the literature related to panel data error components models.

Seemingly, the first authors who dealt with the problem were Mazodier and Trognon (1978). Subsequent contributions include Verbon (1980), Rao *et al.* (1981), Magnus (1982), Baltagi (1988), Baltagi and Griffin (1988), Randolph (1988), Wansbeek (1989), Li and Stengos (1994), Holly and Gardiol (2000), Roy (2002), Phillips (2003), Baltagi *et al.* (2004) and Lejeune (2004).

Within the framework of the classical one-way error components regression model, the issues considered by these papers can be summarized as follows. Both Mazodier and Trognon (1978) and Baltagi and Griffin (1988) are concerned with estimating a model allowing for changing variances of the individual-specific error term across individuals, i.e. they assume that we may write the composite error as $\varepsilon_{it} = \mu_i + \nu_{it}$, $\nu_{it} \sim (0, \sigma_\nu^2)$ while $\mu_i \sim (0, \sigma_{\mu_i}^2)$. Phillips (2003) considers a similar model where heteroscedasticity occurs only through individual-specific variances changing across strata of individuals. Rao *et al.* (1981), Magnus (1982), Baltagi (1988) and Wansbeek (1989) adopt a different specification, allowing for changing variances of the general error term across individuals, i.e. assume that $\nu_{it} \sim (0, \sigma_{\nu_i}^2)$ while $\mu_i \sim (0, \sigma_\mu^2)$. Verbon (1980) is interested in Lagrange Multiplier (LM) testing of the standard normally distributed homoscedastic one-way error components model against the heteroscedastic alternative $\nu_{it} \sim N(0, \sigma_{\nu_i}^2)$ and $\mu_i \sim N(0, \sigma_{\mu_i}^2)$, where $\sigma_{\nu_i}^2$ and $\sigma_{\mu_i}^2$ are, up to a multiplicative constant, identical parametric functions of a vector of time-invariant explanatory variables Z_i, i.e. $\sigma_{\nu_i}^2 = \sigma_\nu^2 \phi(Z_i\gamma)$ and $\sigma_{\mu_i}^2 = \sigma_\mu^2 \phi(Z_i\gamma)$. Baltagi *et al.* (2004) consider a joint LM test of the same null hypothesis but against the more general heteroscedastic alternative $\nu_{it} \sim N(0, \sigma^2_{\nu_{it}})$ and $\mu_i \sim N(0, \sigma^2_{\mu_i})$, where $\sigma^2_{\nu_{it}}$ and $\sigma^2_{\mu_i}$ are, up to a multiplicative constant, possibly different parametric functions of vectors of explanatory variables Z^1_{it} and Z^2_i, i.e. $\sigma^2_{\nu_{it}} = \sigma^2_\nu \phi_\nu(Z^1_{it}\gamma_1)$ and $\sigma^2_{\mu_i} = \sigma^2_\mu \phi_\mu(Z^2_i \gamma_2)$. They further consider "marginal" LM tests of again the same null hypothesis but against the "marginal" heteroscedastic alternatives, on one hand, $\nu_{it} \sim N(0, \sigma^2_{\nu_{it}})$ and $\mu_i \sim N(0, \sigma^2_\mu)$, and on the other hand, $\nu_{it} \sim N(0, \sigma^2_\nu)$ and $\mu_i \sim N(0, \sigma^2_{\mu_i})$. The latter test was previously obtained by Holly and Gardiol (2000). Lejeune (2004) provides a distribution-free joint test and robust one-directional tests of the null hypothesis of no individual effect and heteroscedasticity. These tests allow one to detect, from preliminary (pooled) OLS estimation of the model, the possible simultaneous presence of both individual effects and heteroscedasticity. Randolph (1988) concentrates on supplying an observation-by-observation data transformation for a full heteroscedastic error components model assuming that $\nu_{it} \sim (0, \sigma^2_{\nu_{it}})$ and $\mu_i \sim (0, \sigma^2_{\mu_i})$. Provided that the variances $\sigma^2_{\nu_{it}}$ and $\sigma^2_{\mu_i}$ are known, this transformation allows generalized least squares estimates to be obtained from ordinary least squares. Li and Stengos (1994) deal with adaptive estimation of an error components model supposing heteroscedasticity of unknown form for the general error term, i.e. assume that $\mu_i \sim (0, \sigma^2_\mu)$ while $\nu_{it} \sim (0, \sigma^2_{\nu_{it}})$, where $\sigma^2_{\nu_{it}}$ is a non-parametric function $\phi(Z_{it})$ of a vector of explanatory variables Z_{it}. Likewise, Roy (2002) considers adaptive estimation of a error components model also assuming heteroscedasticity of unknown form, but for the individual-specific error term, i.e. supposes that $\mu_i \sim (0, \sigma^2_{\mu_i})$ while $\nu_{it} \sim (0, \sigma^2_\nu)$. Except Rao *et al.* (1981), Randolph (1988) and Lejeune (2004), all these papers consider balanced panels.

In this paper, we are concerned with estimation and specification testing of a full heteroscedastic one-way error components linear regression model specified in the spirit of Randolph (1988) and Baltagi *et al.* (2004). In short, we assume that the (conditional) variances $\sigma^2_{\nu_{it}}$ and $\sigma^2_{\mu_i}$ are distinct parametric functions of, respectively, vectors of explanatory variables Z^1_{it} and Z^2_i, i.e. $\sigma^2_{\nu_{it}} = \phi_\nu(Z^1_{it}\gamma_1)$ and $\sigma^2_{\mu_i} = \phi_\mu(Z^2_i \gamma_2)$. Further, we treat the model in the context of unbalanced panels. This specification differs from the previously proposed formulations of estimable heteroscedastic error components models as it simultaneously embodies three characteristics. First, heteroscedasticity distinctly applies to both individual-specific and general error components. Second, (non-linear) variance functions are parametrically specified. Finally, the model allows for unbalanced panels.

Explicitly allowing for unbalanced panels is obviously desirable. Indeed, at least for micro-data, incompleteness is rather the rule than the exception. Specifying parametric variance functions is also attractive. First, this strategy avoids incidental parameter (and thus consistency) problems arising from any attempt to model changing variances by grouped heteroscedasticity when the number of individual units is large but the number of observations per individual is small, i.e. in typical microeconomic panel datasets. Second, provided that the functional forms of the variance functions are judiciously chosen, it prevents problems due to estimated variances being negative or zero. Finally, since the variance estimates may have intrinsic values of their own as indicators of the between and within individual heterogeneity, parametric forms are convenient for ease of interpretation.

The heuristic background for allowing heteroscedasticity to distinctly apply to both individual-specific and general error components is the following. In essence, except for the fact that it may be broken down into an individual-specific and a general component, the composite error term in panel data is not different from a cross-section error term. Accordingly, all we said about the possible sources of heteroscedasticity in cross-section may be roughly applied to the panel data composite error term. The only new issue is to assess the plausible origin – between and/or within, i.e. the individual-specific error and/or the general error – of any given cross-section like heteroscedasticity in the composite error term. Clearly, the answer depends upon the situation at hand. When heteroscedasticity arises from differences in size, both error terms may be expected to be heteroscedastic, presumably according to parallel patterns. As a matter of fact, this is implicitly acknowledged whenever a transformation of the dependent variable is used for solving heteroscedasticity problems (the transformation alters the distribution of both error terms). Likewise, if size-related heteroscedasticity still prevails after having transformed the dependent variable, the same should hold. When heteroscedasticity is not directly associated with size, it seems much more difficult to say anything general: depending on the situation, either only one or both error terms may be heteroscedastic, and when both are, their scedastic pattern may further be different. Be that as it may, as a general setting, it thus appears sensible to allow heteroscedasticity to distinctly apply to both individual-specific and general error components.

For estimating our proposed full heteroscedastic one-way error components model, we argue for resorting to a Gaussian pseudo-maximum likelihood of order 2 estimator (Gourieroux *et al.*, 1984; Gourieroux and Monfort, 1993; Bollerslev and Wooldridge, 1992; Wooldridge, 1994). This estimator has indeed numerous nice properties: it is computationally convenient, it allows one to straightforwardly handle unbalanced panels, it is efficient under normality but robust to non-normality, and last but not least, in the present context, it is also robust to possible misspecification of the assumed scedastic structure of the data.

Further, we outline how, taking advantage of the powerful m-testing framework (Newey, 1985; Tauchen, 1985; White, 1987, 1994; Wooldridge, 1990, 1991a, 1991b), the correct specification of the prominent aspects of our proposed model may be tested. We consider potentially useful nested, non-nested, Hausman and information matrix type diagnostic tests of both the mean and the variance specifications. Joined to the Gaussian pseudo-maximum likelihood of order 2 (GPML2) estimator, this set of diagnostic tests provides a complete statistical tool-box for estimating and evaluating the empirical relevance of our proposed model. For Gauss users, an easy-to-use package implementing this complete statistical tool-box may be obtained (free of charge) upon request from the author.

The rest of the paper proceeds as follows. Section 2.2 describes our proposed full heteroscedastic one-way error components model. Section 2.3 considers GPML2 estimation of the model and outlines its asymptotic properties. Section 2.4 deals with specification testing of the model. Section 2.5 provides an empirical illustration of the practical usefulness of our suggested model and estimation and specification testing procedures. Finally, Section 2.6 concludes.

2.2 The model

We consider the following one-way error components linear regression model

$$Y_{it} = X_{it}\beta + \varepsilon_{it}, \quad \varepsilon_{it} = \mu_i + \nu_{it}, \ i = 1, 2, \ldots, n; \ t = 1, 2, \ldots, T_i, \tag{2.1}$$

where Y_{it}, ε_{it}, μ_i and ν_{it} are scalars, X_{it} is a $1 \times k$ vector of strictly exogenous explanatory variables (the first element being a constant) and β is a $k \times 1$ vector of parameters. The index i refers to the individuals and the index t to the (repeated) observations (over time) of each individual i. Each individual i is assumed to be observed a fixed number of times T_i. The unbalanced structure of the panel is supposed to be ignorable in the sense of Wooldridge (1995). The total number of observations is $N = \sum_{i=1}^{n} T_i$. The observations are assumed to be independently (but not necessarily identically) distributed across individuals.

Stacking the T_i observations of each individual i, (2.1) yields the multivariate linear regression model

$$Y_i = X_i\beta + \varepsilon_i, \quad \varepsilon_i = e_{T_i}\mu_i + \nu_i, \ i = 1, 2, \ldots, n, \tag{2.2}$$

where e_{T_i} is a $T_i \times 1$ vector of ones, Y_i, ν_i and ε_i are $T_i \times 1$ vectors, and X_i is a $T_i \times k$ matrix.

Let Z_i^1 denote a $T_i \times l_1$ matrix of strictly exogenous explanatory variables (the first column being a constant), Z_{it}^1 stand for the tth row of Z_i^1, and Z_i^2 be a $1 \times l_2$ vector of strictly exogenous explanatory variables (the first element being again a constant). For all i, t and t', the error terms ν_{it} and μ_i are assumed to satisfy the assumptions

$$E\bigl(\nu_{it} | X_i, Z_i^1, Z_i^2\bigr) = 0, \qquad E\bigl(\mu_i | X_i, Z_i^1, Z_i^2\bigr) = 0, \tag{2.3}$$

$$\begin{aligned} &E\bigl(\nu_{it}\nu_{it'} | X_i, Z_i^1, Z_i^2\bigr) = 0 \quad (t' \neq t), \\ &E\bigl(\mu_i \nu_{it} | X_i, Z_i^1, Z_i^2\bigr) = 0, \end{aligned} \tag{2.4}$$

$$\begin{aligned} &V\bigl(\nu_{it} | X_i, Z_i^1, Z_i^2\bigr) = \sigma_{\nu_{it}}^2 = \phi_\nu\bigl(Z_{it}^1 \gamma_1\bigr) \quad \text{and} \\ &V\bigl(\mu_i | X_i, Z_i^1, Z_i^2\bigr) = \sigma_{\mu_i}^2 = \phi_\mu\bigl(Z_i^2 \gamma_2\bigr), \end{aligned} \tag{2.5}$$

where $\phi_\nu(\cdot)$ and $\phi_\mu(\cdot)$ are (strictly) positive twice continuously differentiable functions while γ_1 and γ_2 are, respectively, $l_1 \times 1$ and $l_2 \times 1$ vectors of parameters which vary independently of each other and independently of β. Hereafter, we will denote by $\gamma = (\gamma_1', \gamma_2')'$ the vector of variance-specific parameters, and $\theta = (\beta', \gamma')'$ will stand for the entire set of parameters.

The regressors appearing in the conditional variances (2.5) may (and usually will) be related to the X_i variables. Different choices are possible for the variance functions $\phi_\nu(\cdot)$ and $\phi_\mu(\cdot)$, see for example Breusch and Pagan (1979) and Harvey (1976). Among them, the multiplicative heteroscedasticity formulation investigated in Harvey (1976) appears particularly attractive. It simply means taking $\phi_\nu(\cdot) = \phi_\mu(\cdot) = \exp(\cdot)$.

Under (2.3)–(2.5), ε_i is easily seen to satisfy

$$\begin{aligned} &E\bigl(\varepsilon_i | X_i, Z_i^1, Z_i^2\bigr) = 0, \quad i = 1, 2, \ldots, n, \\ &V\bigl(\varepsilon_i | X_i, Z_i^1, Z_i^2\bigr) = \Omega_i = \operatorname{diag}\bigl(\phi_\nu\bigl(Z_i^1 \gamma_1\bigr)\bigr) + J_{T_i}\phi_\mu\bigl(Z_i^2 \gamma_2\bigr), \end{aligned} \tag{2.6}$$

where $J_{T_i} = e_{T_i} e_{T_i}'$ is a $T_i \times T_i$ matrix of ones, and, for a $T_i \times 1$ vector x, the functions $\phi_\nu(x)$ and $\phi_\mu(x)$ denote $T_i \times 1$ vectors containing the element-by-element transformations $\phi_\nu(x)$ and $\phi_\mu(x)$ of the elements of x, $\operatorname{diag}(\phi_\nu(x))$ further standing for a diagonal $T_i \times T_i$ matrix containing $\phi_\nu(x)$ as diagonal elements and zeros elsewhere.

The model may thus be written as

$$\begin{aligned} &E\bigl(Y_i | X_i, Z_i^1, Z_i^2\bigr) = X_i \beta, \quad i = 1, 2, \ldots, n, \\ &V\bigl(Y_i | X_i, Z_i^1, Z_i^2\bigr) = \Omega_i = \operatorname{diag}\bigl(\phi_\nu\bigl(Z_i^1 \gamma_1\bigr)\bigr) + J_{T_i}\phi_\mu\bigl(Z_i^2 \gamma_2\bigr). \end{aligned} \tag{2.7}$$

This model obviously contains the standard homoscedastic one-way error components linear regression model as a special case: it is simply obtained by letting the Z_i^1 and Z_i^2 variables only contain an intercept.

In practice, model (2.7) may or may not be correctly specified. It will be correctly specified for the conditional mean if the observations are indeed such that $E(Y_i|X_i, Z_i^1, Z_i^2) = X_i\beta^o$, $i = 1, 2, \ldots, n$, for some true value β^o. Likewise, it will be correctly specified for the conditional variance if the observations are indeed such that $V(Y_i|X_i, Z_i^1, Z_i^2) = \Omega_i^o = \text{diag}(\phi_\nu(Z_i^1\gamma_1^o)) + J_{T_i}\phi_\mu(Z_i^2\gamma_2^o)$, $i = 1, 2, \ldots, n$, for some true-value $\gamma^o = (\gamma_1^{o\prime}, \gamma_2^{o\prime})'$.

2.3 Pseudo-maximum likelihood estimation

The most popular procedure for estimating the standard homoscedastic one-way error components model consists in first estimating the mean parameters of the model by OLS, then in estimating the variance of the error components based on the residuals obtained in the first step, and finally, for efficiency, in re-estimating the mean parameters by feasible generalized least squares (FGLS).

Pursuing a similar multiple-step procedure for estimating our proposed full heteroscedastic model does not appear very attractive. Indeed, if in the standard homoscedastic model it is straightforward to consistently estimate the variance of the error components based on first step regression residuals, it is no longer the case in our proposed full heteroscedastic model: given the general functional forms adopted for the variance functions,[1] no simple – i.e. avoiding non-linear optimization – procedure for consistently estimating the variance parameters appearing in Ω_i seems conceivable.

As non-linear optimization appears unavoidable, we argue for estimating our proposed model by Gaussian pseudo-maximum likelihood of order two (Gourieroux *et al.*, 1984; Gourieroux and Monfort, 1993; Bollerslev and Wooldridge, 1992; Wooldridge, 1994). This GPML2 estimator has numerous attractive properties. First, if it requires non-linear optimization, it is a one-step estimator, simultaneously providing mean and variance parameters estimates. Second, as developed below, while fully efficient if normality holds, it is not only robust to non-normality (i.e. its consistency

[1] The problem would be different if the variance functions were assumed linear. Specifying linear variance functions is however not a good idea as it may result in estimated variances being negative or zero.

does not rely on normality) but also to possible misspecification of the conditional variance (i.e. it remains consistent for the mean parameters even if the assumed scedastic structure of the data is misspecified). Finally, it readily allows one to handle unbalanced panels.

2.3.1 *The GPML2 estimator*

The GPML2 estimator $\hat{\theta}_n = (\hat{\beta}'_n, \hat{\gamma}'_{1_n}, \hat{\gamma}'_{2_n})'$ of model (2.7) is defined as a solution of

$$\operatorname*{Max}_{\theta\in\Theta} L_n(\beta,\gamma_1,\gamma_2) = \frac{1}{n}\sum_{i=1}^{n} L_i\big(Y_i|X_i, Z_i^1, Z_i^2; \beta,\gamma_1,\gamma_2\big), \tag{2.8}$$

where Θ denotes the parameter space and the (conditional) pseudo log-likelihood functions $L_i(Y_i|X_i, Z_i^1, Z_i^2; \beta,\gamma_1,\gamma_2)$ are

$$L_i\big(Y_i|X_i, Z_i^1, Z_i^2; \beta,\gamma_1,\gamma_2\big) = -\frac{T_i}{2}\ln 2\pi - \frac{1}{2}\ln|\Omega_i| - \frac{1}{2}u_i'\Omega_i^{-1}u_i$$

with $u_i = Y_i - X_i\beta$.

Closed-form expressions are available for $|\Omega_i|$ and Ω_i^{-1}. These are given in Appendix A2, where we also provide expressions for the first derivatives, Hessian matrix and expected Hessian matrix of the pseudo log-likelihood function $L_n(\beta,\gamma_1,\gamma_2)$.

If one checks the first-order conditions defining $\hat{\theta}_n$, it is evident that the GPML2 mean-specific estimator $\hat{\beta}_n$ is nothing but a FGLS estimator where the variance parameters appearing in Ω_i are jointly estimated. Additionally, the GPML2 variance-specific estimator $\hat{\gamma}_n = (\hat{\gamma}'_{1_n}, \hat{\gamma}'_{2_n})'$ may be interpreted as a weighted non-linear least squares estimator in the multivariate non-linear regression model $\operatorname{vec}(u_iu_i') = \operatorname{vec}\Omega_i(\gamma_1,\gamma_2) +$ residuals, $i = 1, 2, \ldots, n$, where the errors u_i and the weights $\Gamma_i^{-1} = (\Omega_i^{-1}\otimes\Omega_i^{-1})$ are likewise jointly estimated.

Practical guidelines for computing the GPML2 estimator $\hat{\theta}_n$, including a numerical algorithm and starting values, are discussed in Appendix B2.

2.3.2 *Asymptotic properties of the GPML2 estimator*

Beyond its computational convenience and its ability to readily handle unbalanced panel, the most attractive feature of the GPML2 estimator is its statistical properties, namely its potential efficiency and its robustness.

Obviously, when the model is correctly specified for both the conditional mean and the conditional variance and when in addition normality also holds, the GPML2 estimator is just a standard maximum likelihood estimator. According to standard maximum likelihood theory, we then have that $\hat{\theta}_n$ is consistent and asymptotically normal,

$$\hat{\theta}_n \xrightarrow{p} \theta^o \quad \text{and} \quad \sqrt{n}\big(\hat{\theta}_n - \theta^o\big) \approx \mathcal{N}\big(0, \overline{C}_n^o\big),$$
$$\text{as } n \to \infty \ (T_i \text{ bounded})$$

with an asymptotic covariance matrix given by

$$\overline{C}_n^o = \begin{bmatrix} -\overline{A}_{\beta\beta}^{-1} & 0 \\ 0 & -\overline{A}_{\gamma\gamma}^{-1} \end{bmatrix},$$

where

$$\overline{A}_{\beta\beta} = \frac{1}{n}\sum_{i=1}^{n} E\big[\underline{h}_i^{\beta\beta}\big]_{\theta=\theta^o}, \qquad \overline{A}_{\gamma\gamma} = \frac{1}{n}\sum_{i=1}^{n} E\big[\underline{h}_i^{\gamma\gamma}\big]_{\theta=\theta^o},$$
$$\underline{h}_i^{\gamma\gamma} = \begin{bmatrix} \underline{h}_i^{\gamma_1\gamma_1} & \underline{h}_i^{\gamma_1\gamma_2} \\ \underline{h}_i^{\gamma_2\gamma_1} & \underline{h}_i^{\gamma_2\gamma_2} \end{bmatrix}$$

and $\underline{h}_i^{\beta\beta}$ and $\underline{h}_i^{\gamma\gamma}$ refer to the expected Hessian of L_i and are defined in Appendix A2.

In this favorable situation, the GPML2 estimator is fully efficient, both for the mean and the variance parameters. However, since in practice normality may at best be expected to only very approximately hold, this result must essentially be viewed as a benchmark result.

As for all pseudo-maximum likelihood estimators, the distributional normality assumption underlying the GPML2 estimator is purely nominal. As a matter of fact, according to second order Gaussian pseudo-maximum likelihood theory (Gourieroux *et al.*, 1984; Gourieroux and Monfort, 1993; Bollerslev and Wooldridge, 1992; Wooldridge, 1994), if the model is correctly specified for the conditional mean and the conditional variance but normality does not hold, we still have that $\hat{\theta}_n$ is consistent and asymptotically normal,

$$\hat{\theta}_n \xrightarrow{p} \theta^o \quad \text{and} \quad \sqrt{n}\big(\hat{\theta}_n - \theta^o\big) \approx \mathcal{N}\big(0, C_n^o\big),$$
$$\text{as } n \to \infty \ (T_i \text{ bounded})$$

but with a more complicated asymptotic covariance matrix given by

$$C_n^o = \begin{bmatrix} -\overline{A}_{\beta\beta}^{-1} & \overline{A}_{\beta\beta}^{-1} B_{\beta\gamma} \overline{A}_{\gamma\gamma}^{-1} \\ \overline{A}_{\gamma\gamma}^{-1} B_{\gamma\beta} \overline{A}_{\beta\beta}^{-1} & \overline{A}_{\gamma\gamma}^{-1} B_{\gamma\gamma} \overline{A}_{\gamma\gamma}^{-1} \end{bmatrix},$$

where

$$B_{\beta\gamma} = \frac{1}{n}\sum_{i=1}^{n} E\big[s_i^{\beta} s_i^{\gamma\prime}\big]_{\theta=\theta^o} = B'_{\gamma\beta}, \qquad B_{\gamma\gamma} = \frac{1}{n}\sum_{i=1}^{n} E\big[s_i^{\gamma} s_i^{\gamma\prime}\big]_{\theta=\theta^o},$$
$$s_i^{\gamma} = \big(s_i^{\gamma_1\prime}, s_i^{\gamma_2\prime}\big)',$$

and s_i^β and s_i^γ refer to the first derivatives of L_i and are again defined in Appendix A2.

Note that non-normality does not affect the asymptotic covariance matrix of the GPML2 mean-specific estimator $\hat{\beta}_n$. It is still given by $-\overline{A}_{\beta\beta}^{-1}$, which, since $\hat{\beta}_n$ is in fact nothing but a FGLS estimator, is actually equal to the asymptotic covariance matrix of the usual FGLS estimator (implemented using any consistent estimator of the variance parameters appearing in Ω_i). Of course, in this situation, the GPML2 estimator is no longer fully efficient. It is clearly not efficient regarding the variance parameters. Regarding the mean parameters, as FGLS, it is however still efficient in a semi-parametric sense.[2]

Besides being robust to non-normality, the GPML2 estimator has an additional nice property in that it is also robust to conditional variance misspecification, i.e. to misspecification of the assumed scedastic structure of the data. Since the GPML2 mean-specific estimator $\hat{\beta}_n$ is a FGLS estimator, this should not be surprising.[3] According to Lejeune (1998), if the model is correctly specified for the conditional mean but misspecified for the conditional variance, it indeed turns out that $\hat{\beta}_n$ is still consistent for its true value β^o while $\hat{\gamma}_n$ is now consistent for some pseudo-true value $\gamma_n^* = (\gamma_{1_n}^{*\prime}, \gamma_{2_n}^{*\prime})'$,

$$\hat{\beta}_n \xrightarrow{p} \beta^o \quad \text{and} \quad \hat{\gamma}_n - \gamma_n^* \xrightarrow{p} 0, \quad \text{as } n \to \infty \ (T_i \text{ bounded})$$

and that $\hat{\theta}_n$ remains jointly asymptotically normal

$$\sqrt{n}\big(\hat{\theta}_n - \theta_n^{o*}\big) \approx \mathcal{N}\big(0, C_n^{o*}\big),$$
$$\text{as } n \to \infty \ (T_i \text{ bounded}), \ \text{where } \theta_n^{o*} = \big(\beta^{o\prime}, \gamma_n^{*\prime}\big)'$$

with an asymptotic covariance matrix given by

$$C_n^{o*} = \begin{bmatrix} A_{\beta\beta}^{-1} B_{\beta\beta} A_{\beta\beta}^{-1} & A_{\beta\beta}^{-1} B_{\beta\gamma} A_{\gamma\gamma}^{-1} \\ A_{\gamma\gamma}^{-1} B_{\gamma\beta} A_{\beta\beta}^{-1} & A_{\gamma\gamma}^{-1} \ddot{B}_{\gamma\gamma} A_{\gamma\gamma}^{-1} \end{bmatrix},$$

where

$$A_{\beta\beta} = \frac{1}{n}\sum_{i=1}^{n} E\big[h_i^{\beta\beta}\big]_{\theta=\theta_n^{o*}}, \qquad A_{\gamma\gamma} = \frac{1}{n}\sum_{i=1}^{n} E\big[h_i^{\gamma\gamma}\big]_{\theta=\theta_n^{o*}},$$

[2] The asymptotic covariance matrix of $\hat{\beta}_n$ attains the well-known semi-parametric efficiency bound (Chamberlain, 1987; Newey, 1990, 1993; Wooldridge, 1994) associated with optimal GMM estimation based on the first-order conditional moments of the data.

[3] It is well-known that conditional variance misspecification does not affect the consistency of the FGLS estimator.

$$h_i^{\gamma\gamma} = \begin{bmatrix} h_i^{\gamma_1\gamma_1} & h_i^{\gamma_1\gamma_2} \\ h_i^{\gamma_2\gamma_1} & h_i^{\gamma_2\gamma_2} \end{bmatrix},$$

$$B_{\beta\beta} = \frac{1}{n}\sum_{i=1}^{n} E\big[s_i^{\beta} s_i^{\beta\prime}\big]_{\theta=\theta_n^{o*}}, \qquad B_{\beta\gamma} = \frac{1}{n}\sum_{i=1}^{n} E\big[s_i^{\beta} s_i^{\gamma\prime}\big]_{\theta=\theta_n^{o*}} = B'_{\gamma\beta},$$

$$\ddot{B}_{\gamma\gamma} = \frac{1}{n}\sum_{i=1}^{n} E\big[s_i^{\gamma} s_i^{\gamma\prime}\big]_{\theta=\theta_n^{o*}} - U_{\gamma\gamma},$$

$$U_{\gamma\gamma} = \frac{1}{n}\sum_{i=1}^{n} E\big[s_i^{\gamma}\big]_{\theta=\theta_n^{o*}} E\big[s_i^{\gamma}\big]'_{\theta=\theta_n^{o*}}$$

and $h_i^{\beta\beta}$ and $h_i^{\gamma\gamma}$ refer to the Hessian of L_i and are again defined in Appendix A2.

Of course, in this latter situation, the GPML2 mean-specific estimator $\hat{\beta}_n$ is no longer efficient. However, as its asymptotic covariance matrix $A_{\beta\beta}^{-1} B_{\beta\beta} A_{\beta\beta}^{-1}$ collapses to the semi-parametric efficiency bound $-\overline{A}_{\beta\beta}^{-1}$ outlined above when the conditional variance is correctly specified, we may intuitively expect that the more the specified conditional variance is close to the actual scedastic structure of the data, the more the covariance matrix of $\hat{\beta}_n$ will be close to this lower bound, i.e. $\hat{\beta}_n$ will be close to semi-parametric efficiency. From a empirical point of view, this in particular implies that it makes sense to consider using our proposed full heteroscedastic model, even if possibly misspecified, whenever the homoscedasticity assumption of the standard one-way error components model does not appear to hold: some efficiency benefits may indeed generally be expected from taking into account even approximately the actual scedastic structure of the data.

In practical applications, the extent to which our assumed full heteroscedastic model is actually correctly specified is of course *a priori* unknown. This may nevertheless be checked through diagnostic tests, as discussed in Section 2.4 below. Once this is done, a consistent estimate of the asymptotic covariance matrix of the estimated parameters may then be straightforwardly computed by taking, as usual, the empirical counterpart of the relevant theoretical asymptotic covariance matrix.[4] There

[4] For example, a consistent estimate of the asymptotic covariance matrix $A_{\beta\beta}^{-1} B_{\beta\beta} A_{\beta\beta}^{-1}$ of the GPML2 mean-specific estimator $\hat{\beta}_n$ under correct conditional mean specification but conditional variance misspecification may be computed as $\widehat{A}_{\beta\beta}^{-1} \widehat{B}_{\beta\beta} \widehat{A}_{\beta\beta}^{-1}$, where $\widehat{A}_{\beta\beta} = \frac{1}{n}\sum_{i=1}^{n} \hat{h}_i^{\beta\beta}$, $\widehat{B}_{\beta\beta} = \frac{1}{n}\sum_{i=1}^{n} \hat{s}_i^{\beta} \hat{s}_i^{\beta\prime}$ and the superscript '^' denotes quantities evaluated at $\hat{\theta}_n$.

is one exception however: due to the term $U_{\gamma\gamma}$, unless the observations are IID and the panel dataset is balanced (in which case $U_{\gamma\gamma} = 0$), a consistent estimate of the asymptotic covariance matrix $A_{\gamma\gamma}^{-1}\ddot{B}_{\gamma\gamma}A_{\gamma\gamma}^{-1}$ of the GPML2 variance-specific estimator $\hat{\gamma}_n$ under correct conditional mean specification but conditional variance misspecification may in general not be obtained. A consistent estimate of an upper bound of this asymptotic covariance matrix, upper bound given by $A_{\gamma\gamma}^{-1}B_{\gamma\gamma}A_{\gamma\gamma}^{-1}$ where $B_{\gamma\gamma} = \frac{1}{n}\sum_{i=1}^{n} E[s_i^{\gamma} s_i^{\gamma\prime}]_{\theta=\theta_n^{o*}}$, may nevertheless be computed in the usual way. Interestingly, based on this estimated upper bound, a conservative – i.e. with asymptotic true size necessarily inferior to its specified nominal size – (joint) Wald test of the null hypothesis that the non-intercept parameters of γ_1 and γ_2 are zero may then be validly performed. In other words, a valid conservative test which checks that, as assumed, the observations indeed exhibit some heteroscedasticity-like pattern related to the Z_i^1 and Z_i^2 explanatory variables may then readily be carried out, and this is regardless of possible conditional variance misspecification.

2.4 Specification testing

The GPML2 estimator of model (2.7) always delivers a consistent estimate of the mean parameters if the model is correctly specified for the conditional mean, and consistent estimates of both the mean and variance parameters if the model is correctly specified for both the conditional mean and the conditional variance. But nothing *a priori* guarantees that the model is indeed correctly specified.

Hereafter, we outline how, taking advantage of the powerful m-testing framework (Newey, 1985; Tauchen, 1985; White, 1987, 1994; Wooldridge, 1990, 1991a, 1991b), the conditional mean and the conditional variance specification of our proposed full heteroscedastic one-way error components model may be checked. We first consider conditional mean diagnostic tests, and then conditional variance diagnostic tests.

2.4.1 Conditional mean diagnostic tests

Having estimated our proposed model (2.7), the first thing to consider is to check its conditional mean specification. Testing the null hypothesis that the conditional mean is correctly specified means testing

$$\mathrm{H}_0^m\colon\ E\big(Y_i|X_i, Z_i^1, Z_i^2\big) = X_i\beta^o, \quad \text{for some } \beta^o,\ i = 1, 2, \ldots, n.$$

Following White (1987, 1994), Wooldridge (1990, 1991a, 1991b) and Lejeune (1998), based on the GPML2 estimator $\hat{\theta}_n$, H_0^m may efficiently be tested by checking, for appropriate choices of $T_i \times q$ indicator matrices $\widehat{W}_i^m$ (which may depend on the conditioning variables (X_i, Z_i^1, Z_i^2) as well as on additional estimated nuisance parameters), that $q \times 1$ misspecification indicators of the form

$$\widehat{\varPhi}_n^m = \frac{1}{n}\sum_{i=1}^{n} \widehat{W}_i^{m\prime}\widehat{\varOmega}_i^{-1}\hat{u}_i \tag{2.9}$$

are not significantly different from zero.

Given the assumed statistical setup, a relevant statistic for checking that $\widehat{\varPhi}_n^m$ is not significantly different from zero is given by the asymptotic chi-squared statistic[5]

$$\begin{aligned}\mathcal{M}_n^m = {} & \left(\sum_{i=1}^{n} \widehat{W}_i^{m\prime}\widehat{\varOmega}_i^{-1}\hat{u}_i\right)' \\ & \times \left(\sum_{i=1}^{n}\big(\widehat{W}_i^m - X_i\widehat{P}^m\big)'\widehat{\varOmega}_i^{-1}\hat{u}_i\hat{u}_i'\widehat{\varOmega}_i^{-1}\big(\widehat{W}_i^m - X_i\widehat{P}^m\big)\right)^{-1} \\ & \times \left(\sum_{i=1}^{n} \widehat{W}_i^{m\prime}\widehat{\varOmega}_i^{-1}\hat{u}_i\right) \xrightarrow{d} \chi^2(q),\end{aligned}$$

where

$$\widehat{P}^m = \left(\sum_{i=1}^{n} X_i'\widehat{\varOmega}_i^{-1}X_i\right)^{-1}\sum_{i=1}^{n} X_i'\widehat{\varOmega}_i^{-1}\widehat{W}_i^m.$$

By suitably choosing the $T_i \times q$ indicator matrices $\widehat{W}_i^m$ in (2.9), as detailed below, H_0^m may be tested against nested alternatives, non-nested alternatives, or without resorting to explicit alternatives through Hausman and information matrix type tests.

A prominent characteristic of all conditional mean diagnostic tests implemented through the $\mathcal{M}_n^m$ statistic is that they yield valid tests of H_0^m regardless of whether or not the assumed scedastic pattern of the data is correct and whether or not normality holds. Consequently, since they do not rely on assumptions other than H_0^m itself, a rejection may always be unambiguously attributed to a failure of H_0^m to hold. Interestingly, another important characteristic of diagnostic tests implemented through $\mathcal{M}_n^m$ is

[5] Note that $\mathcal{M}_n^m$ may in practice be computed as n minus the residual sum of squares $(= nR_u^2$, R_u^2 being the uncentered R-squared) of the artificial OLS regression $1 = [\hat{u}_i'\widehat{\varOmega}_i^{-1}(\widehat{W}_i^m - X_i\widehat{P}^m)]b +$ residuals, $i = 1, 2, \ldots, n$.

that they will have optimal properties if the conditional variance is actually correctly specified and normality holds.

Following Wooldridge (1990, 1991a, 1991b) and Lejeune (1998), for testing H_0^m against a nested alternative of the form

$$\begin{aligned}&\mathrm{H}_1^m\colon\ E\big(Y_i|X_i, Z_i^1, Z_i^2\big) = m_i^a\big(X_i, Z_i^1, Z_i^2, \beta^o, \alpha^o\big),\\ &\quad\text{for some } \big(\beta^{o\prime}, \alpha^{o\prime}\big)',\ i = 1, 2, \ldots, n,\end{aligned}$$

where $m_i^a(X_i, Z_i^1, Z_i^2, \beta, \alpha)$ denotes some alternative conditional mean specification such that for some value $\alpha = c$ of the $q \times 1$ vector of additional parameters α we have

$$m_i^a\big(X_i, Z_i^1, Z_i^2, \beta, c\big) = X_i\beta, \quad i = 1, 2, \ldots, n,$$

the appropriate choice of $\widehat{W}_i^m$ is given by

$$\widehat{W}_i^m = \frac{\partial m_i^a(X_i, Z_i^1, Z_i^2, \hat{\beta}_n, c)}{\partial \alpha'}.$$

When the considered alternative conditional mean specification takes the simple linear form

$$m_i^a\big(X_i, Z_i^1, Z_i^2, \beta, \alpha\big) = X_i\beta + G_i\alpha, \quad i = 1, 2, \ldots, n,$$

where G_i is a $T_i \times q$ matrix of variables which are functions of the set of conditioning variables $CV_i \equiv (X_i, Z_i^1, Z_i^2)$, $\widehat{W}_i^m$ is simply equal to G_i and the test corresponds to a standard variable addition test. We may for example check in this way the linearity of the assumed conditional mean by setting G_i equal to (some of) the squares and/or the cross-products of (some of) the X_i variables.

On the other hand, for testing H_0^m against a non-nested alternative such as

$$\begin{aligned}&\mathrm{H}_1^m\colon\ E\big(Y_i|X_i, Z_i^1, Z_i^2\big) = g_i^a\big(X_i, Z_i^1, Z_i^2, \delta^o\big),\\ &\quad\text{for some } \delta^o,\ i = 1, 2, \ldots, n,\end{aligned}$$

where $g_i^a(X_i, Z_i^1, Z_i^2, \delta)$ denotes some alternative conditional mean specification which does not contain the null conditional mean specification $X_i\beta$ as a special case and δ is a vector of parameters, an appropriate choice of $\widehat{W}_i^m$ is given by

$$\widehat{W}_i^m = g_i^a\big(X_i, Z_i^1, Z_i^2, \hat{\delta}_n\big) - X_i\hat{\beta}_n,$$

where $\hat{\delta}_n$ is any consistent estimator of δ^o under H_1^m. This yields a Davidson and MacKinnon (1981) type test of a non-nested alternative. Because obvious choices of $g_i^a(\cdot)$ are in practice rarely available, this kind of test of H_0^m is unlikely to be routinely performed. It may however be useful in some situations.

By construction, diagnostic tests against nested or non-nested alternatives have power against the specific alternative they consider, but may be expected to have limited power against other (if weakly related) alternatives. General purpose diagnostic tests with expected power against a broader range of alternatives are provided by Hausman and information matrix type tests.

One of the equivalent forms of the popular Hausman specification test of the standard homoscedastic one-way error components model is based on comparing the (non-intercept) FGLS and OLS estimators of β^o (see for example Baltagi, 1995). This strongly suggests considering a generalized (i.e. allowing for any choice of S and robust to conditional variance misspecification) Hausman type test of H_0^m based on checking, for some chosen selection matrix S, the closeness to zero of the misspecification indicator

$$\widehat{\Phi}_n^m = S\big(\hat{\beta}_n - \hat{\beta}_n^{\mathrm{OLS}}\big).$$

Following the lines of White (1994) and Lejeune (1998), a test that is asymptotically equivalent to checking the above misspecification indicator is obtained by setting

$$\widehat{W}_i^m = \widehat{\Omega}_i X_i \widehat{Q}^{-1} S',$$

where $\widehat{Q} = \sum_{i=1}^n X_i' X_i$. As is the case with the standard textbook Hausman test (to which it is asymptotically equivalent under standard textbook homoscedasticity conditions), this test will have power against any alternative H_1^m for which $\hat{\beta}_n$ and $\hat{\beta}_n^{\mathrm{OLS}}$ converge to different pseudo-true values. Note by the way that, contrary to the standard textbook case, heteroscedasticity (and incompleteness) usually allows one to include all β parameters as part of this Hausman test without yielding a singular statistic.

On the other hand, following again the lines of White (1994) and Lejeune (1998), an information matrix type test of H_0^m may be based on checking, for some chosen selection matrix S, the closeness to zero of the misspecification indicator

$$\widehat{\Phi}_n^m = S\frac{1}{n}\sum_{i=1}^n \operatorname{vec} \hat{h}_i^{\beta\gamma}, \quad h_i^{\beta\gamma} = \big[\, h_i^{\beta\gamma_1} \quad h_i^{\beta\gamma_2} \,\big],$$

where $h_i^{\beta\gamma}$ refers to cross-derivatives of L_i and is defined in Appendix A2. Such a test essentially involves checking the block diagonality between mean and variance parameters of the expected Hessian matrix of the GPML2 estimator, which must hold under correct conditional mean specification (regardless of the correctness of the conditional variance specification). It is obtained by setting

$$\widehat{W}_i^m = \widehat{F}_i S',$$

where

$$\widehat{F}_i = \left[\frac{\partial \widehat{\Omega}_i}{\partial \gamma_1^1} \widehat{\Omega}_i^{-1} X_i \cdots \frac{\partial \widehat{\Omega}_i}{\partial \gamma_1^{l_1}} \widehat{\Omega}_i^{-1} X_i \; \frac{\partial \widehat{\Omega}_i}{\partial \gamma_2^1} \widehat{\Omega}_i^{-1} X_i \cdots \frac{\partial \widehat{\Omega}_i}{\partial \gamma_2^{l_2}} \widehat{\Omega}_i^{-1} X_i \right]$$

and $\frac{\partial \Omega_i}{\partial \gamma_p^r}$ $(p = 1, 2)$ is again defined in Appendix A2. This test, which will have power against any alternative H_1^m for which the block diagonality of the expected Hessian matrix the GPML2 estimator fails, is a quite natural complement to the above Hausman test for testing H_0^m without resorting to explicit alternatives. Note that if the multiplicative heteroscedasticity formulation is adopted for both $\phi_\nu(\cdot)$ and $\phi_\mu(\cdot)$, one of the two matrix elements $\frac{\partial \widehat{\Omega}_i}{\partial \gamma_1^1} \widehat{\Omega}_i^{-1} X_i$ and $\frac{\partial \widehat{\Omega}_i}{\partial \gamma_2^1} \widehat{\Omega}_i^{-1} X_i$ of $\widehat{F}_i$ is redundant (yielding a singular statistic for S being set to an identity matrix) and must thus be discarded.

When a test against a specific nested or non-nested alternative rejects the null hypothesis H_0^m, it is natural to then consider modifying the originally assumed conditional mean specification in the direction of the considered alternative. When a Hausman or information matrix type test rejects H_0^m, the way that one should react is less obvious and depends on the situation at hand. In all cases, considering further diagnostic tests against various nested or non-nested alternatives should help one to identify the source(s) of rejection of H_0^m.

To conclude this brief review of conditional mean diagnostic m-tests, we make one additional remark. In empirical practice, it is not unusual for one to test the null model against an explicit alternative which includes variables which are not functions of the original set of conditioning variables $CV_i \equiv (X_i, Z_i^1, Z_i^2)$. This does not modify the way in which testing against explicit alternatives is implemented. It is however important to be aware that, in such a case, we are no longer only testing the null H_0^m but instead the null $\mathrm{H}_0^{m\prime}$: H_0^m holds and $E(Y_i | X_i, Z_i^1, Z_i^2, \underline{G_i}) = E(Y_i | X_i, Z_i^1, Z_i^2)$, $i = 1, 2, \ldots, n$, where $\underline{G_i}$ denotes the variables which are not functions of CV_i. In other words, we are jointly testing that H_0^m holds and that the additional $\underline{G_i}$ variables are irrelevant as conditioning variables for the expectation of Y_i. We thus must be careful in interpreting such a specification test given that H_0^m might well hold while $\mathrm{H}_0^{m\prime}$ does not.

2.4.2 Conditional variance diagnostic tests

Having tested – and if needed adjusted – the conditional mean specification of the model, we may then check its conditional variance specification. Testing the null hypothesis that the conditional variance is correctly specified entails testing the null

$$\mathrm{H}_0^v\colon \begin{cases} \mathrm{H}_0^m \text{ holds and, for some } \gamma^o, \\ V\big(Y_i|X_i, Z_i^1, Z_i^2\big) = \operatorname{diag}\big(\phi_\nu\big(Z_i^1\gamma_1^o\big)\big) + J_{T_i}\phi_\mu\big(Z_i^2\gamma_2^o\big), \\ \quad i = 1, 2, \ldots, n. \end{cases}$$

Note that H_0^v embodies H_0^m: there is indeed no way to test the conditional variance specification without simultaneously assuming that the conditional mean is correctly specified. This is however not a real problem since, using the above diagnostic tests, the conditional mean specification may in a first step be checked without having to assume correct conditional variance specification.

Following again White (1987, 1994), Wooldridge (1990, 1991a, 1991b) and Lejeune (1998), based on the GPML2 estimator $\hat{\theta}_n$, H_0^v may efficiently be tested by checking, for appropriate choices of $T_i^2 \times q$ indicator matrices $\widehat{W}_i^v$ (which may depend on the conditioning variables (X_i, Z_i^1, Z_i^2) as well as on additional estimated nuisance parameters), that $q \times 1$ misspecification indicators which similarly are of the form

$$\widehat{\Phi}_n^v = \frac{1}{n}\sum_{i=1}^{n} \widehat{W}_i^{v\prime}\widehat{\Gamma}_i^{-1}\hat{v}_i, \tag{2.10}$$

where

$$\widehat{\Gamma}_i^{-1} = \big(\widehat{\Omega}_i^{-1} \otimes \widehat{\Omega}_i^{-1}\big) \quad \text{and} \quad \hat{v}_i = \operatorname{vec}(\hat{u}_i\hat{u}_i' - \widehat{\Omega}_i),$$

are not significantly different from zero.

Given the assumed statistical setup, a relevant statistic for checking that $\widehat{\Phi}_n^v$ is not significantly different from zero is given by the asymptotic chi-squared statistic[6]

$$\begin{aligned} \mathcal{M}_n^v = {} & \left(\sum_{i=1}^{n} \widehat{W}_i^{v\prime}\widehat{\Gamma}_i^{-1}\hat{v}_i\right)' \\ & \times \left(\sum_{i=1}^{n}\left(\widehat{W}_i^v - \frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma'}\widehat{P}^v\right)' \right. \\ & \left. \times\, \widehat{\Gamma}_i^{-1}\hat{v}_i\hat{v}_i'\widehat{\Gamma}_i^{-1}\left(\widehat{W}_i^v - \frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma'}\widehat{P}^v\right)\right)^{-1} \end{aligned}$$

[6] Note that $\mathcal{M}_n^v$ may in practice be computed as n minus the residual sum of squares $(= nR_u^2$, R_u^2 being the uncentered R-squared) of the artificial OLS regression $1 = [\hat{v}_i'\widehat{\Gamma}_i^{-1}(\widehat{W}_i^v - \frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma'}\widehat{P}^v)]b + \text{residuals}$, $i = 1, 2, \ldots, n$.

$$\times \left(\sum_{i=1}^{n} \widehat{W}_i^{v\prime} \widehat{\Gamma}_i^{-1} \hat{v}_i \right) \xrightarrow{d} \chi^2(q),$$

where

$$\frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma'} = \left[\frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma_1'} \quad \frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma_2'} \right],$$

$$\widehat{P}^v = \left(\sum_{i=1}^{n} \left(\frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma'} \right)' \widehat{\Gamma}_i^{-1} \frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma'} \right)^{-1} \sum_{i=1}^{n} \left(\frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma'} \right)' \widehat{\Gamma}_i^{-1} \widehat{W}_i^v$$

and $\frac{\partial \operatorname{vec} \Omega_i}{\partial \gamma'}$ is defined in Appendix A2.

As was the case with the conditional mean diagnostic tests, by suitably choosing the $T_i^2 \times q$ indicator matrices $\widehat{W}_i^v$ in (2.10), as detailed below, H_0^v may be tested against nested alternatives, non-nested alternatives, or without resorting to explicit alternatives through Hausman and information matrix type tests.

A prominent characteristic of all conditional variance diagnostic tests implemented through the $\mathcal{M}_n^v$ statistic is that they yield valid tests of H_0^v whether or not normality holds. Consequently, since they do not rely on assumptions other than H_0^v itself, a rejection may always be unambiguously attributed to a failure of H_0^v to hold. Further, given the nested nature of H_0^m and H_0^v and the robustness to possible conditional variance misspecification of the diagnostic tests of H_0^m, if no misspecification has been detected by conditional mean diagnostic tests, a rejection of H_0^v may then sensibly be attributed to conditional variance misspecification: situations where conditional variance diagnostic tests detect a misspecification in the mean which has not been detected by conditional mean diagnostic tests are indeed likely to be rare in practice. Interestingly, another important characteristic of diagnostic tests implemented through the $\mathcal{M}_n^v$ is that they will have optimal properties if normality actually holds.

Following White (1994), Wooldridge (1990, 1991a, 1991b) and Lejeune (1998), for testing H_0^v against a nested alternative of the form

$$\mathrm{H}_1^v\colon \begin{cases} \mathrm{H}_0^m \text{ holds and, for some } (\gamma^{o\prime}, \alpha^{o\prime})', \\ V(Y_i | X_i, Z_i^1, Z_i^2) = \Omega_i^a(X_i, Z_i^1, Z_i^2, \gamma^o, \alpha^o), \quad i = 1, 2, \ldots, n, \end{cases}$$

where $\Omega_i^a(X_i, Z_i^1, Z_i^2, \gamma, \alpha)$ denotes some alternative conditional variance specification such that for some value $\alpha = c$ of the $q \times 1$ vector of additional parameters α we have

$$\Omega_i^a(X_i, Z_i^1, Z_i^2, \gamma, c) = \operatorname{diag}(\phi_\nu(Z_i^1 \gamma_1)) + J_{T_i} \phi_\mu(Z_i^2 \gamma_2),$$
$$i = 1, 2, \ldots, n,$$

the appropriate choice of $\widehat{W}_i^v$ is given by

$$\widehat{W}_i^v = \frac{\partial \operatorname{vec} \Omega_i^a(X_i, Z_i^1, Z_i^2, \hat{\gamma}_n, c)}{\partial \alpha'}.$$

If the considered nested alternative takes the simple semi-linear form

$$\begin{aligned}\Omega_i^a\big(X_i, Z_i^1, Z_i^2, \gamma, \alpha\big) = {} & \operatorname{diag}\big(\phi_\nu\big(Z_i^1\gamma_1 + G_i^1\alpha_1\big)\big) \\ & + J_{T_i}\phi_\mu\big(Z_i^2\gamma_2 + G_i^2\alpha_2\big),\end{aligned}$$

where $\alpha = (\alpha_1', \alpha_2')'$ and G_i^1 and G_i^2 are respectively $T_i \times q_1$ matrices and $1 \times q_2$ vectors ($q_1 + q_2 = q$) of variables which are functions of the set of conditioning variables $CV_i \equiv (X_i, Z_i^1, Z_i^2)$, the test corresponds to a variable addition test and $\widehat{W}_i^v$ is equal to

$$\widehat{W}_i^v = \big[\, \widehat{W}_i^{v1} \quad \widehat{W}_i^{v2} \,\big]$$

with

$$\begin{aligned}\widehat{W}_i^{v1} &= \operatorname{diag}\big(\operatorname{vec}\big(\operatorname{diag}\big(\phi_\nu'\big(Z_i^1\hat{\gamma}_{1_n}\big)\big)\big)\big)\big(G_i^1 \otimes e_{T_i}\big) \\ &= \sum_{r=1}^{q_1} \operatorname{vec}\big(\operatorname{diag}\big(\phi_\nu'\big(Z_i^1\hat{\gamma}_{1_n}\big) \odot G_i^{1^r}\big)\big)e_{q_1}^{r'},\end{aligned}$$

$$\widehat{W}_i^{v2} = \phi_\mu'\big(Z_i^2\hat{\gamma}_{2_n}\big)\operatorname{vec}(J_{n_i})G_i^2 = \sum_{r=1}^{q_2} \operatorname{vec}\big(\phi_\mu'\big(Z_i^2\hat{\gamma}_{2_n}\big)G_i^{2^r}J_{T_i}\big)e_{q_2}^{r'},$$

where $\phi_\nu'(\cdot)$ and $\phi_\mu'(\cdot)$ stand for the first derivatives of $\phi_\nu(\cdot)$ and $\phi_\mu(\cdot)$, $G_i^{1^r}$ and $G_i^{2^r}$ denote the rth column of respectively G_i^1 and G_i^2, $e_{q_1}^r$ and $e_{q_2}^r$ are respectively $q_1 \times 1$ and $q_2 \times 1$ vectors with a one in the rth place and zeros elsewhere, and $\odot$ stands for the Hadamard product, i.e. an element-by-element multiplication. As for the conditional mean, we may for example check in this way the semi-linearity of the assumed conditional variance by setting G_i^1 and G_i^2 equal to (some of) the squares and/or the cross-products of (some of) the Z_i^1 and Z_i^2 variables.

On the other hand, for testing H_0^v against a non-nested alternative such as

$$\mathrm{H}_1^v\colon \begin{cases} \mathrm{H}_0^m \text{ holds and, for some } \delta^o, \\ V\big(Y_i|X_i, Z_i^1, Z_i^2\big) = \Sigma_i^a\big(X_i, Z_i^1, Z_i^2, \delta^o\big), \quad i = 1, 2, \ldots, n, \end{cases}$$

where $\Sigma_i^a(X_i, Z_i^1, Z_i^2, \delta)$ denotes some alternative conditional variance specification which does not contain the null conditional variance specification Ω_i as a special case and δ is a vector of parameters, appropriate choices of $\widehat{W}_i^v$ are given by

$$\widehat{W}_i^v = \operatorname{vec}\big(\widehat{\Sigma}_i^a - \widehat{\Omega}_i\big) \tag{2.11}$$

and

$$\widehat{W}_i^v = \operatorname{vec}\big(\widehat{\Omega}_i \widehat{\Sigma}_i^{a^{-1}} \widehat{\Omega}_i - \widehat{\Omega}_i\big), \tag{2.12}$$

where $\widehat{\Sigma}_i^a = \Sigma_i^a(X_i, Z_i^1, Z_i^2, \hat{\delta}_n)$ and $\hat{\delta}_n$ is any consistent estimator of δ^o under H_1^v. The first possible choice (2.11) of $\widehat{W}_i^v$ yields a Davidson and MacKinnon (1981) type test of a non-nested alternative while the second one (2.12) corresponds to a Cox (1961, 1962) type test of a non-nested alternative. It seems that the Cox-like form of the test is generally more powerful than the Davidson-like form. Be that as it may, such tests may for example be used for checking the chosen variance functions $\phi_\nu(\cdot)$ and $\phi_\mu(\cdot)$ against some other possible functional forms, or more generally for checking the assumed heteroscedastic model against any other non-nested specification for the scedastic structure of the data.

As was the case in our discussion of conditional mean testing, when a test against a specific nested or non-nested alternative rejects the null hypothesis H_0^v, it is natural for one to consider modifying the originally assumed conditional variance specification in the direction of the considered alternative. Likewise, in both the nested and non-nested cases, the way to perform the tests is unchanged if the alternative includes variables which are not functions of the original set of conditioning variables $CV_i \equiv (X_i, Z_i^1, Z_i^2)$. But similarly, the tested null hypothesis is modified. It here takes the form $\mathrm{H}_0^{v\prime}$: H_0^v holds and, both $E(Y_i|X_i, Z_i^1, Z_i^2, \underline{G}_i) = E(Y_i|X_i, Z_i^1, Z_i^2)$ and $V(Y_i|X_i, Z_i^1, Z_i^2, \underline{G}_i) = V(Y_i|X_i, Z_i^1, Z_i^2)$, $i = 1, 2, \ldots, n$, where $\underline{G}_i$ denotes the variables which are not functions of CV_i. In other words, besides H_0^v, $\mathrm{H}_0^{v\prime}$ further assumes that the additional variables $\underline{G}_i$ are irrelevant as conditioning variables for the variance but also for the expectation of Y_i.

Beside tests against nested and non-nested alternatives, general purpose diagnostic tests with expected power against a broader range of alternatives may be performed through Hausman and information matrix type tests.

Testing H_0^v through a Hausman type test requires one to choose a consistent estimator of γ^o alternative to $\hat{\gamma}_n$. As already suggested, the GPML2 estimator $\hat{\gamma}_n$ may be shown to be asymptotically equivalent to the weighted non-linear least squares (NLS) estimator with weights $\{\widetilde{\Gamma}_i^{-1}\}$ of the multivariate non-linear regression $\operatorname{vec}(\tilde{u}_i\tilde{u}_i') = \operatorname{vec}(\operatorname{diag}(\phi_\nu(Z_i^1\gamma_1)) + J_{T_i}\phi_\mu(Z_i^2\gamma_2)) +$ residuals, $i = 1, 2, \ldots, n$, where the superscript '$\sim$' denotes quantities evaluated at any preliminary consistent estimator of β^o

and γ^o. A straightforward and natural alternative to it is hence to use the standard (i.e. unweighted) NLS estimator, say $\underline{\hat{\gamma}}_n$, of the same non-linear regression. Accordingly, a relevant Hausman type test of H_0^v may be obtained by checking, for some chosen selection matrix S, the closeness to zero of the misspecification indicator

$$\widehat{\Phi}_n^v = S(\hat{\gamma}_n - \underline{\hat{\gamma}}_n).$$

Following the lines of White (1994) and Lejeune (1998), a test asymptotically equivalent to checking the above misspecification indicator is obtained by setting

$$\widehat{W}_i^v = \widehat{\Gamma}_i \frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma'} \widehat{Q}^{-1} S',$$

where $\widehat{Q} = \sum_{i=1}^n (\frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma'})' \frac{\partial \operatorname{vec} \widehat{\Omega}_i}{\partial \gamma'}$. As with all Hausman type tests, this test will have power against any alternative H_1^v for which $\hat{\gamma}_n$ and $\underline{\hat{\gamma}}_n$ converge to different pseudo-true values.

On the other hand, following again the lines of White (1994) and Lejeune (1998), an information matrix type test of H_0^v may be based on checking, for some chosen selection matrix S which at least removes its otherwise obvious redundant elements, the closeness to zero of the misspecification indicator

$$\widehat{\Phi}_n^v = S \frac{1}{n} \sum_{i=1}^n \operatorname{vec}\bigl(\hat{s}_i^\beta \hat{s}_i^{\beta\prime} + \underline{\hat{h}}_i^{\beta\beta}\bigr).$$

Such a test basically means checking the information matrix equality $B_{\beta\beta} = -\overline{A}_{\beta\beta}$ for the mean parameters, which must hold under correct conditional mean and conditional variance specification. It is obtained by setting

$$\widehat{W}_t^v = (X_i \otimes X_i) S'.$$

This latter way of testing H_0^v without resorting to explicit alternatives, which seems generally more powerful than the above Hausman type test, will clearly have power against any alternative H_1^v for which the mean parameters information matrix equality fails.

As in conditional mean testing, when a Hausman or information matrix type test rejects H_0^v, the way to react is not obvious and depends on the situation at hand. But in all cases, considering further diagnostic tests against various nested or non-nested alternatives should likewise help to identify the source(s) of rejection of H_0^v.

2.5 An empirical illustration

We hereafter illustrate the potential usefulness of our proposed full heteroscedastic model and its accompanying robust inferential methods through an empirical example which involves estimating and testing at an inter-sectorial level the correctness of the specification of a transcendental logarithmic (translog) production model for a sample of 824 French firms observed over the period 1979–1988. As we will see, the results of this exercise suggest (a) that, as argued in Baltagi and Griffin (1988), heteroscedasticity-related problems are likely to be present when estimating this kind of production model, (b) that our proposed full heteroscedastic model and its accompanying robust inferential methods offer a sensible, although imperfect, way to deal with it, and (c) that a judicious use of the set of proposed specification tests allows one to obtain very informative insights regarding the empirical correctness of this simple production model.

2.5.1 Data and model

The data originally came from a panel dataset constructed by the "Marchés et Stratégie d'Entreprises" division of INSEE. It contains 5 201 observations and involves an unbalanced panel of 824 French firms from 9 sectors[7] of the NAP 15 Classification observed over the period 1979–1988.[8] Available data include the value added (va) of the firms deflated by an NAP 40 sector-specific price index (base: 1980), their stock of capital (k) and their labor force (l). The stock of capital has been constructed by INSEE and the labor force is the number of workers expressed in full-time units.

As is usual in this kind of dataset, the variability of the observations essentially lies in the between (across individuals) dimension and is very important: the number of workers ranges from 19 to almost 32 000 and the capital intensity (k/l) varies from a factor of 1 to more than 320. Globally, large firms are over-represented.

For this dataset, we considered estimating and testing the following full heteroscedastic one-way error components translog production function model:

$$\begin{aligned} V_{it} = \beta_{(sc\times t)} + \beta_k K_{it} + \beta_l L_{it} + \beta_{kk} K_{it}^2 + \beta_{ll} L_{it}^2 + \beta_{kl} K_{it} L_{it} \\ + \mu_i + \nu_{it} \end{aligned} \tag{2.13}$$

[7] Agricultural and food industries, energy production and distribution, intermediate goods industries, equipment goods industries, consumption goods industries, construction and civil engineering, trade, transport and telecommunications, and market services.

[8] I wish to thank Patrick Sevestre for giving me the opportunity to use this dataset.

with

$$\sigma^2_{\nu_{it}} = \exp\bigl(\gamma_1^c + \gamma_1^k K_{it} + \gamma_1^l L_{it}\bigr), \tag{2.14}$$

$$\sigma^2_{\mu_i} = \exp\bigl(\gamma_2^c + \gamma_2^k \overline{K}_i + \gamma_2^l \overline{L}_i\bigr), \tag{2.15}$$

where

$$V_{it} = \ln va_{it}, \qquad K_{it} = (\ln k_{it} - \ln k^*), \qquad L_{it} = (\ln l_{it} - \ln l^*),$$

$$\overline{K}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} K_{it} \quad \text{and} \quad \overline{L}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} L_{it}.$$

The subscript '$(sc \times t)$' attached to the intercept parameter $\beta_{(sc\times t)}$ means that we actually let the intercept be sectorial and time-period specific. The model thus contains 90 dummies (9 sectors × 10 periods). This allows for sector-specific productivity growth patterns.

The explanatory variables are centered so that the estimated values of β_k and β_l reported below may directly be interpreted as the elasticities of the value added with respect to capital and labor at $k = k^*$ and $l = l^*$. We set k^* and l^* at their entire sample means.

For both the individual-specific and general error variance functions, we adopted Harvey's (1976) multiplicative heteroscedasticity formulation. In the general error variance function, the explanatory variables are simply taken as the (log of the) capital and labor inputs. Taking the individual mean values of the (log of the) capital and labor inputs as explanatory variables in the individual-specific variance function is mainly a pragmatic choice. It appears sensible as far as the observations variability prominently lies in the between dimension. Be that as it may, these choices allow the variances to change according to both size and input ratios.

2.5.2 *Estimation and specification testing*

The results of GPML2 estimation of model (2.13)–(2.15) are reported in Table 2.1.[9] As it seems natural when first estimating the model, the covariance matrix of the parameters was first computed supposing correct conditional mean specification but possibly misspecified conditional variance, i.e. as the empirical counterpart of C_n^{o*}, or more precisely as the empirical counterpart of C_n^{o*} for the mean parameters and as the empirical counterpart of the outlined upper bound (thus allowing Wald conservative tests) of C_n^{o*} for the variance parameters (see Section 2.3.2). The standard errors reported in Table 2.1 are derived from this first estimated covariance matrix.

[9] For conciseness, the dummy parameter estimates are not reproduced.

Table 2.1. GPML2 estimates and diagnostic tests

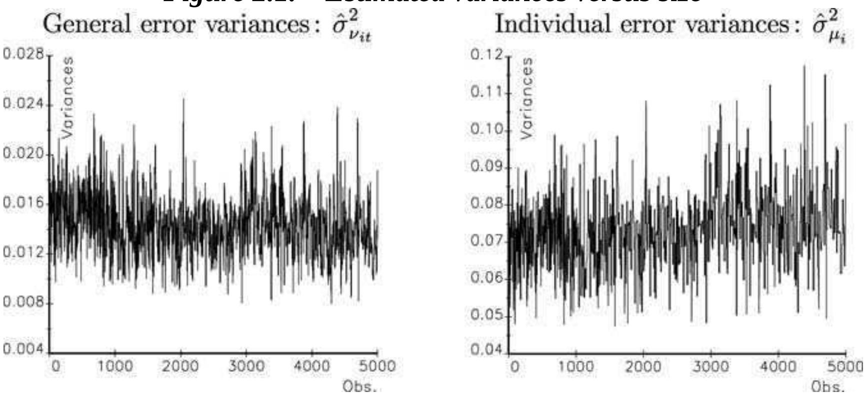
Variable	Coefficient	Std. error*	t-ratio	P-value
K	0.2487	0.0188	13.26	0.0000
L	0.7367	0.0244	30.21	0.0000
K^2	0.0547	0.0072	7.58	0.0000
L^2	0.0572	0.0132	4.35	0.0000
KL	−0.1137	0.0176	−6.48	0.0000
$\sigma^2_{\nu_{it}} = \exp(\cdot)$				
const.	−4.1997	0.0541	−77.65	0.0000
K	0.1870	0.0582	3.21	0.0013
L	−0.2482	0.0849	−2.92	0.0035
$\sigma^2_{\mu_i} = \exp(\cdot)$				
const.	−2.5213	0.0732	−34.43	0.0000
$\overline{K}$	0.1676	0.0610	2.74	0.0060
$\overline{L}$	−0.1709	0.0799	−2.14	0.0325

	Stat.	D.f.	P-value
Conditional mean tests			
(1) Hausman	5.9	5	0.3180
(2) Information matrix	33.7	25	0.1141
(3) H_1: non-neutral TP	8.4	2	0.0146
(4) H_1: third power	2.8	4	0.5961
(5) H_1: time heterogeneity	57.1	45	0.1064
(6) H_1: sectorial heterogeneity	41.0	40	0.4249
Conditional variance tests			
(7) Hausman	18.4	6	0.0052
(8) Information matrix	45.6	15	0.0001
(9) H_1: second power	2.2	6	0.9015
(10) H_1: sectorial heterogeneity	98.6	48	0.0000

*Standard errors computed assuming correct conditional mean specification but possibly misspecified conditional variance.

As is apparent from Table 2.1 and confirmed when formally performing a (conservative) Wald test of the null hypothesis that the non-intercept parameters of both individual-specific and general variance functions are zero (P-value of the test: 0.0008), it appears that heteroscedasticity-like patterns are effectively present in both the individual-specific and general errors of the model. In both cases, heteroscedasticity seems to be related to input ratios: more capital intensive firms tend to achieve more heterogeneous outputs both in the between and within dimensions relative to the more labor intensive firms. The captured heteroscedasticity does not however seem to be notably related to size. Figure 2.1 portrays this latter

Figure 2.1. Estimated variances versus size



General error variances: $\hat{\sigma}^2_{\nu_{it}}$

Individual error variances: $\hat{\sigma}^2_{\mu_i}$

point. In this figure, estimated general error and individual-specific error variances are graphed against the observations sorted in ascending order according to individual means of the fitted dependent variable and, within each individual, according to the values of the fitted dependent variable itself.

Neither of these plots reveal notable links between variances and size. They do however outline two other points. First, variations in the observed inputs ratios imply variations in the estimated variances – identified by the difference between the lower and upper levels of the estimated variances – of more than a factor 2. Second, the estimated individual-specific variances are roughly 5–6 times higher than the estimated general error variances.

Having estimated the model, we next checked the correctness of its specification, considering first its conditional mean specification. To this end, we performed both Hausman and information matrix type tests and tests against nested alternatives. For the record, Hausman and information matrix type tests may be viewed as general purpose diagnostic tests allowing one to in particular detect unforeseen forms of misspecification, while tests against nested alternatives constitute a standard device for detecting a priori well-defined and plausible forms of misspecification.

In the present case, we considered a Hausman test based on comparing the GPML2 and OLS estimators of all mean parameters (excepted the dummies) and an information matrix test based on checking the closeness to zero of the sub-block of the Hessian corresponding to the cross-derivatives between the non-intercept mean parameters and all variance parameters (except for the intercept of the individual-specific variance function, to avoid singularity (cf. Section 2.4.1)). On the other hand, we considered tests against nested alternatives checking for possible non-neutral technical progress (the alternative model including as additional variables the interactions between a trend and the first-order terms of the translog function[10]), for a possible more general functional form (the alternative model including terms of third power[11] as additional variables to the null translog specification), for possible time heterogeneity (the alternative model allowing for the non-intercept mean parameters to be time-period specific), and finally for possible sectorial heterogeneity (the alternative model allowing for the non-intercept mean parameters to be sector-specific).

Table 2.1 reports the results obtained from the computation of these conditional mean diagnostic tests.[12] As may be seen, it appears that the conditional mean does not exhibit patent misspecification. The only statistic which indicates some possible deviation from correct specification is the one of test (3). Its P-value is however not really worrying: from a formal point of view, according to a standard Bonferroni approach, for rejecting at 5% the null hypothesis that the conditional mean is correctly specified, we “need” that at least one of the 6 separate tests rejects the null at 0.83% ($0.05/6 \simeq 0.0083$). Viewed in a less formal way, it is normal to find that some statistics (moderately) deviate when multiplying the number of diagnostic tests. The model may thus sensibly be viewed as a satisfactory statistical representation – on which for example productivity growth measurements could be based – of the available data for the conditional mean.

Taking correct conditional mean specification of the model for granted, we then examined its conditional variance. To this end, as for the conditional mean, we performed general purpose Hausman and information matrix type tests and tests against nested alternatives. Practically, we considered a Hausman test based on comparing the GPML2 and (unweighted) NLS estimators of all variance parameters and an information matrix test based on checking the closeness to zero of the non-redundant elements of the sub-block of the information matrix equality associated with the non-intercept mean parameters. On the other hand, we considered tests against nested alternatives checking for a possible more general functional form

[10] Non-neutral technical progress is typically modeled by considering a trend, a trend-squared and interaction terms between the trend and the first-order terms of the translog function as additional inputs. The trend and trend-squared terms being already captured by the set of dummies, it thus remains to test for the interaction terms between a trend and the first-order terms of the translog function.

[11] I.e. K^3, L^3, KL^2 and K^2L.

[12] Note that none of these diagnostic tests involves variables which are not a function of the original set of conditioning variables (i.e. K, L, sector dummies and time dummies). The null hypothesis of these tests is thus never more than H_0^m itself (cf. Section 2.4.1).

(the alternative model specifying both the individual-specific and general error variances as (the exponential of) translog functions instead of Cobb–Douglas like functions) and for possible sectorial heterogeneity (the alternative model allowing for all variance parameters to be sector-specific).

Before examining the results of these tests,[13] note that the fact of finding no patent misspecification in the conditional mean supports the validity of the (conservative) standard errors of the variance parameter estimates reported in Table 2.1. These standard errors – and further the result of the outlined formal (conservative) Wald test of the null hypothesis that the non-intercept parameters of the individual-specific and general variance functions are zero – undoubtedly indicate that a heteroscedasticity-like pattern is effectively present in the errors of the model. However, according to the conditional variance tests reported in the same table, the assumed specification for this heteroscedasticity-like pattern turns out to be seriously misspecified. Test (9) suggests that relaxing the functional form would not really help. On the other hand, test (10) points out that a problem of sectorial heterogeneity might be involved.

To shed light on the latter point as well as to gauge the sensibility of the conditional mean estimates and diagnostic tests to the specification of the conditional variance, Table 2.2 reports GPML2 estimates and diagnostic tests – the same tests as above – of an extension of model (2.13)–(2.15), where both the individual-specific and the general error variance parameters are allowed to be sector-specific.

As may be seen from Table 2.2, the obtained mean parameter estimates are not very different from those obtained under the assumption of identical variances across sectors (cf. Table 2.1). For conciseness, the variance parameter estimates are not reported. But, as expected, they unambiguously confirm both that a heteroscedasticity-like pattern related to input ratios is present, and that this heteroscedasticity-like pattern is indeed sector-specific.

The diagnostic tests reported in Table 2.2 corroborate our result that the conditional mean of the model does not exhibit patent misspecification. However, they also show that allowing for sector-specific variance functions did not solve our misspecification problem in the conditional variance. How to fix this misspecification does not appear to be a trivial exercise.

Note nevertheless that, even if misspecified, these sector-specific variance functions are not useless. Comparing the standard errors of the mean

[13] Note again that none of these diagnostic tests involves variables which are not a function of the original set of conditioning variables. The null hypothesis of these tests is thus again never more than H_0^v itself (cf. Section 2.4.2).

Table 2.2. GPML2 estimates and diagnostic tests with sector-specific conditional variances

Variable	Coefficient	Std. error*	t-ratio	P-value
K	0.2455	0.0169	14.54	0.0000
L	0.7519	0.0210	35.77	0.0000
K^2	0.0557	0.0062	9.03	0.0000
L^2	0.0639	0.0101	6.29	0.0000
KL	−0.1165	0.0148	−7.87	0.0000
		Stat.	D.f.	P-value
Conditional mean tests				
(1) Hausman		6.5	5	0.2579
(2) Information matrix		38.7	25	0.0396
(3) H_1: non-neutral TP		3.9	2	0.1446
(4) H_1: third power		3.3	4	0.5061
(5) H_1: time heterogeneity		55.6	45	0.1341
(6) H_1: sectorial heterogeneity		36.0	40	0.6505
Conditional variance tests				
(7) Hausman		72.1	50	0.0221
(8) Information matrix		52.8	15	0.0000

*Standard errors computed assuming correct conditional mean specification but possibly misspecified conditional variance.

parameters reported in Tables 2.1 and 2.2, it may indeed be seen that allowing for this more flexible conditional variance specification has entailed (moderate) efficiency gains: the reduction of the standard errors ranges from -10.1% to -23.4%. This illustrates that, as argued in Section 2.3.2, a misspecified conditional variance may get efficiency benefits – for estimation but also testing of the conditional mean – from taking into account even approximately the actual scedastic structure of the data.

2.6 Conclusion

This paper proposed an extension of the standard one-way error components model allowing for heteroscedasticity in both the individual-specific and the general error terms, as well as for unbalanced panel. On the grounds of its computational convenience, its ability to straightforwardly handle unbalanced panels, its potential efficiency, its robustness to non-normality and its robustness to possible misspecification of the assumed scedastic structure of the data, we argued for estimating this model by Gaussian pseudo-maximum likelihood of order two. We further reviewed how, taking advantage of the powerful m-testing framework, the correct specification of the prominent aspects of the assumed full heteroscedastic model may be tested. We finally illustrated the practical relevance of our proposed model and estimation and diagnostic testing procedures through an empirical example.

To conclude, note that, since our proposed model contains as a special case the standard one-way error components model (just let the Z_i^1 and Z_i^2 variables only contain an intercept), our proposed integrated statistical tool-box, for which an easy-to-use Gauss package is available upon request from the author, may actually also be used for estimating and checking the specification of this standard model. On the other hand, remark that, following the lines of this paper, our proposed integrated statistical tool-box may readily be adapted to handle a more general model, for example allowing for a nonlinear (instead of linear) specification in the conditional mean and/or any fully nonlinear (instead of semi-linear) specification in the conditional variance.

Acknowledgements

I wish to thank Laurence Broze, Patrick Sevestre, Alain Trognon, Jean-Pierre Urbain, James Lindsey, Badi Baltagi, Tim Coelli, Muriel Fonder, Arie Preminger and two anonymous referees for useful comments and suggestions on earlier versions of this paper. Errors, of course, remain my sole responsibility.

Appendix A2

Closed-form expressions for $|\Omega_i|$ and Ω_i^{-1} are given by

$$|\Omega_i| = (b_i)^{T_i}|C_i|\big(1 + \operatorname{tr} C_i^{-1}\big) = \left(\prod_{t=1}^{T_i} a_{it}\right)(1 + e'_{T_i}\bar{c}_i),$$

$$\begin{aligned}\Omega_i^{-1} &= \frac{1}{b_i}\left(C_i^{-1} - \frac{1}{1 + \operatorname{tr} C_i^{-1}}\big(C_i^{-1} J_{T_i} C_i^{-1}\big)\right)\\ &= \operatorname{diag}(\bar{a}_i) - \frac{1}{b_i(1 + e'_{T_i}\bar{c}_i)}\bar{c}_i\bar{c}_i',\end{aligned}$$

where

$$b_i = \phi_\mu\big(Z_i^2\gamma_2\big), \qquad c_i = \frac{1}{b_i}\phi_\nu\big(Z_i^1\gamma_1\big), \qquad a_i = \phi_\nu\big(Z_i^1\gamma_1\big),$$

$$C_i = \operatorname{diag}(c_i), \qquad \bar{c}_i = e_{T_i} \div c_i, \qquad \bar{a}_i = e_{T_i} \div a_i,$$

a_{it} being the tth element of a_i and $\div$ indicating an element-by-element division. Note that according to this notation, $\Omega_i = b_i(C_i + J_{T_i})$.

Following Magnus (1978, 1988), the first derivatives of $L_n(\beta, \gamma_1, \gamma_2)$ may be written

$$\frac{\partial L_n}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} s_i^\theta, \quad s_i^\theta = \begin{bmatrix} s_i^\beta \\ s_i^{\gamma_1} \\ s_i^{\gamma_2} \end{bmatrix},$$

with

$$s_i^\beta = X_i'\Omega_i^{-1}u_i, \tag{A2.1}$$

$$\begin{aligned} s_i^{\gamma_p} &= \frac{1}{2}\left(\frac{\partial \operatorname{vec}\Omega_i}{\partial \gamma_p'}\right)'\big(\Omega_i^{-1}\otimes\Omega_i^{-1}\big)\operatorname{vec}(u_iu_i' - \Omega_i) \quad (p = 1, 2) \qquad \text{(A2.2)} \\ &= \frac{1}{2}\sum_{r=1}^{l_p}\operatorname{tr}\left(\Omega_i^{-1}\frac{\partial\Omega_i}{\partial\gamma_p^r}\Omega_i^{-1}(u_iu_i' - \Omega_i)\right)e_{l_p}^r, \end{aligned}$$

where $e_{l_p}^r$ is a $l_p \times 1$ vector with a one in the rth place and zeros elsewhere, i.e. the rth column of a $l_p \times l_p$ identity matrix, γ_p^r is the rth component of γ_p, and the derivatives of $\operatorname{vec}\Omega_i$ with respect to γ_p' $(p = 1, 2)$ are

$$\frac{\partial \operatorname{vec}\Omega_i}{\partial\gamma_1'} = \operatorname{diag}\Big(\operatorname{vec}\Big(\operatorname{diag}\Big(\phi_\nu'\big(Z_i^1\gamma_1\big)\Big)\Big)\Big)\big(Z_i^1 \otimes e_{T_i}\big), \tag{A2.3}$$

$$\frac{\partial \operatorname{vec}\Omega_i}{\partial\gamma_2'} = \phi_\mu'\big(Z_i^2\gamma_2\big)\operatorname{vec}(J_{n_i})Z_i^2 = \phi_\mu'\big(Z_i^2\gamma_2\big)(e_{T_i}\otimes e_{T_i})Z_i^2 \tag{A2.4}$$

while the derivatives of Ω_i with respect to γ_p^r $(p = 1, 2)$ are

$$\begin{aligned} \frac{\partial\Omega_i}{\partial\gamma_1^r} &= \operatorname{diag}\big(\phi_\nu'\big(Z_i^1\gamma_1\big)\odot Z_i^{1^r}\big) \quad \text{and} \\ \frac{\partial\Omega_i}{\partial\gamma_2^r} &= \phi_\mu'\big(Z_i^2\gamma_2\big)Z_i^{2^r}J_{T_i}, \end{aligned} \tag{A2.5}$$

where $\phi_\nu'(\cdot)$ and $\phi_\mu'(\cdot)$ denote the first derivatives of $\phi_\nu(\cdot)$ and $\phi_\mu(\cdot)$, $Z_i^{1^r}$ is the rth column of the matrix of explanatory variables Z_i^1, $\odot$ stands for the Hadamard product, i.e. an element-by-element multiplication, and $Z_i^{2^r}$ is the rth column of the row vector of explanatory variables Z_i^2. Note that if the multiplicative heteroscedasticity formulation is adopted for both $\phi_\nu(\cdot)$ and $\phi_\mu(\cdot)$, then, in (A2.3)–(A2.5), $\phi'_\nu(\cdot)$ and $\phi'_\mu(\cdot)$ are simply equal to $\exp(\cdot)$.

Following again Magnus (1978, 1988), the Hessian matrix of $L_n(\beta, \gamma_1, \gamma_2)$ may be written

$$\frac{\partial^2 L_n}{\partial\theta\partial\theta'} = \frac{1}{n}\sum_{i=1}^{n} h_i^{\theta\theta}, \quad h_i^{\theta\theta} = \begin{bmatrix} h_i^{\beta\beta} & h_i^{\beta\gamma_1} & h_i^{\beta\gamma_2} \\ h_i^{\gamma_1\beta} & h_i^{\gamma_1\gamma_1} & h_i^{\gamma_1\gamma_2} \\ h_i^{\gamma_2\beta} & h_i^{\gamma_2\gamma_1} & h_i^{\gamma_2\gamma_2} \end{bmatrix},$$

with

$$h_i^{\beta\beta} = \frac{\partial s_i^{\beta}}{\partial\beta'} = -X_i'\Omega_i^{-1}X_i, \tag{A2.6}$$

$$\begin{aligned} h_i^{\beta\gamma_p} &= \frac{\partial s_i^{\beta}}{\partial\gamma_p'} = \left(\frac{\partial s_i^{\gamma_p}}{\partial\beta'}\right)' = h_i^{\gamma_p\beta'} \quad (p = 1, 2) \\ &= -\left(u_i'\Omega_i^{-1} \otimes X_i'\Omega_i^{-1}\right)\frac{\partial \operatorname{vec}\Omega_i}{\partial\gamma_p'} \\ &= -\sum_{r=1}^{l_p}\left(X_i'\Omega_i^{-1}\frac{\partial\Omega_i}{\partial\gamma_p^r}\Omega_i^{-1}u_i\right)e_{l_p}^{r\prime}, \end{aligned} \tag{A2.7}$$

$$\begin{aligned} h_i^{\gamma_p\gamma_q} &= \frac{\partial s_i^{\gamma_p}}{\partial\gamma_q'} = \left(\frac{\partial s_i^{\gamma_q}}{\partial\gamma_p'}\right)' = h_i^{\gamma_q\gamma_p\prime} \quad (p = 1, 2;\ q = 1, 2) \\ &= -\frac{1}{2}\left(\frac{\partial \operatorname{vec}\Omega_i}{\partial\gamma_p'}\right)'\left(\Omega_i^{-1} \otimes \Omega_i^{-1}\right)\frac{\partial \operatorname{vec}\Omega_i}{\partial\gamma_q'} \\ &\quad - \frac{1}{2}\left(\left(\operatorname{vec}(u_iu_i' - \Omega_i)\right)' \otimes I_{l_p}\right)\Upsilon_i^{\gamma_p\gamma_q} \\ &= -\frac{1}{2}\sum_{r=1}^{l_p}\sum_{s=1}^{l_q}\left(\operatorname{tr}\left(\Omega_i^{-1}\frac{\partial\Omega_i}{\partial\gamma_p^r}\Omega_i^{-1}\frac{\partial\Omega_i}{\partial\gamma_q^s}\right)\right. \\ &\quad \left. + \operatorname{tr}\left((u_iu_i' - \Omega_i)\frac{\partial^2\Omega_i^{-1}}{\partial\gamma_p^r\partial\gamma_q^s}\right)\right)e_{l_p}^{r}e_{l_q}^{s\prime}, \end{aligned} \tag{A2.8}$$

where I_{l_p} is a $l_p \times l_p$ identity matrix,

$$\Upsilon_i^{\gamma_p\gamma_q} = \frac{\partial \operatorname{vec}\left(\frac{\partial \operatorname{vec}\Omega_i^{-1}}{\partial\gamma_p'}\right)'}{\partial\gamma_q'} = \sum_{r=1}^{l_p}\sum_{s=1}^{l_q}\operatorname{vec}\left(\frac{\partial^2\Omega_i^{-1}}{\partial\gamma_p^r\partial\gamma_q^s}\right) \otimes \left(e_{l_p}^{r}e_{l_q}^{s\prime}\right),$$

i.e. a $T_i^2 l_p \times l_q$ matrix,

$$\frac{\partial^2 \Omega_i^{-1}}{\partial \gamma_p^r \partial \gamma_q^s} = \Omega_i^{-1}\left(2\frac{\partial \Omega_i}{\partial \gamma_p^r}\Omega_i^{-1}\frac{\partial \Omega_i}{\partial \gamma_q^s} - \frac{\partial^2 \Omega_i}{\partial \gamma_p^r \partial \gamma_q^s}\right)\Omega_i^{-1}$$

and the needed derivatives not yet given are

$$\begin{aligned}
\frac{\partial^2 \Omega_i}{\partial \gamma_1^r \partial \gamma_1^s} &= \operatorname{diag}\big(\phi_\nu''\big(Z_i^1\gamma_1\big) \odot Z_i^{1^r} \odot Z_i^{1^s}\big),\\
\frac{\partial^2 \Omega_i}{\partial \gamma_1^r \partial \gamma_2^s} &= 0 = \frac{\partial^2 \Omega_i}{\partial \gamma_2^r \partial \gamma_1^s},\\
\frac{\partial^2 \Omega_i}{\partial \gamma_2^r \partial \gamma_2^s} &= \phi_\mu''\big(Z_i^2\gamma_2\big)Z_i^{2^r} Z_i^{2^s} J_{T_i},
\end{aligned}$$

where $\phi_\nu''(\cdot)$ and $\phi_\mu''(\cdot)$ denote the second derivatives of $\phi_\nu(\cdot)$ and $\phi_\mu(\cdot)$. If the multiplicative heteroscedasticity formulation is adopted for both $\phi_\nu(\cdot)$ and $\phi_\mu(\cdot)$, $\phi_\nu''(\cdot)$ and $\phi_\mu''(\cdot)$ are again simply equal to $\exp(\cdot)$.

Under conditional mean and conditional variance correct specification, we have $E(u_i^o|X_i, Z_i^1, Z_i^2) = 0$ and $E((u_i^o u_i^{o\prime} - \Omega_i^o)|X_i, Z_i^1, Z_i^2) = 0$, so that using the law of iterated expectation it is easily checked that the expected Hessian matrix of $L_n(\beta, \gamma_1, \gamma_2)$ may be written

$$E\left[\frac{\partial^2 L_n}{\partial\theta\partial\theta'}\right]_{\theta=\theta^o} = \frac{1}{n}\sum_{i=1}^{n} E\big[h_i^{\theta\theta}\big]_{\theta=\theta^o} = \frac{1}{n}\sum_{i=1}^{n} E\big[\underline{h}_i^{\theta\theta}\big]_{\theta=\theta^o},$$

where

$$\underline{h}_i^{\theta\theta} = \begin{bmatrix} \underline{h}_i^{\beta\beta} & 0 & 0 \\ 0 & \underline{h}_i^{\gamma_1\gamma_1} & \underline{h}_i^{\gamma_1\gamma_2} \\ 0 & \underline{h}_i^{\gamma_2\gamma_1} & \underline{h}_i^{\gamma_2\gamma_2} \end{bmatrix}$$

and

$$\underline{h}_i^{\beta\beta} = h_i^{\beta\beta} = -X_i'\Omega_i^{-1}X_i, \tag{A2.9}$$

$$\begin{aligned}
\underline{h}_i^{\gamma_p\gamma_q} &= \underline{h}_i^{\gamma_q\gamma_p\prime} \quad (p = 1, 2;\ q = 1, 2) \qquad\qquad (A2.10)\\
&= -\frac{1}{2}\left(\frac{\partial \operatorname{vec}\Omega_i}{\partial \gamma_p'}\right)'\big(\Omega_i^{-1} \otimes \Omega_i^{-1}\big)\frac{\partial \operatorname{vec}\Omega_i}{\partial \gamma_q'}\\
&= -\frac{1}{2}\sum_{r=1}^{l_p}\sum_{s=1}^{l_q} \operatorname{tr}\left(\Omega_i^{-1}\frac{\partial \Omega_i}{\partial \gamma_p^r}\Omega_i^{-1}\frac{\partial \Omega_i}{\partial \gamma_q^s}\right)e_{l_p}^r e_{l_q}^{s\prime}.
\end{aligned}$$

Note that contrary to the Hessian which depends on first and second derivatives, the expected Hessian is block-diagonal (between mean and variance parameters) and only depends on first derivatives.

Appendix B2

For Gaussian maximum likelihood estimation of the standard (homoscedastic) one-way error components model, Breusch (1987) suggests an iterated GLS procedure. Although applicable in very general situations (see Magnus, 1978), in the present case it is not very attractive since it implies at each step the (numerical) resolution of a set of non-linear equations defined by the first-order conditions $\frac{\partial L_n}{\partial \gamma_p} = 0$ $(p = 1, 2)$.

As alternatives, we can use either a Newton or quasi-Newton (secant methods) algorithm. While the former requires the computation of the first and second derivatives, the latter (for example, the so-called Davidson–Fletcher–Powell and Broyden–Fletcher–Goldfard–Shanno methods) requires only the computation of the first derivatives (see Quandt, 1983). In the present case, a variant of the Newton method appears particularly appealing, namely the scoring method. This variant simply involves substituting the Hessian $\frac{\partial^2 L_n}{\partial \theta \partial \theta'} = \frac{1}{n} \sum_{i=1}^{n} h_i^{\theta\theta}$ used in the Newton algorithm by the empirical counterpart of its expectation under conditional mean and conditional variance correct specification, i.e. by $\frac{1}{n} \sum_{i=1}^{n} \underline{h}_i^{\theta\theta}$. As noted above in Appendix A2, the latter is considerably simpler: it is block-diagonal and only involves first derivatives. It will be a good approximation of the Hessian if the model is correctly specified and θ is not too far from θ^o. According to our experience, even under quite severe misspecification, provided that all quantities are analytically computed, the scoring method generally converges in less time (more computation time per iteration but fewer iterations) than the secant methods. Further, since the empirical expected Hessian is always negative semidefinite, it is numerically stable.

A sensible set of starting values for the above algorithm may be computed by proceeding as follows:

1. Obtain the $\underline{\hat{\beta}}$ and $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_i, \ldots, \hat{\alpha}_n)$ OLS estimates of the dummy variables model $Y_i = \alpha_i + \underline{X}_i \underline{\beta} +$ residuals $(i = 1, 2, \ldots, n)$, where $\underline{X}_i$ is the same as X_i except its dropped first column. At this stage, $\underline{\hat{\beta}}$ and the mean of the $\hat{\alpha}_i$, i.e. $\bar{\alpha} = \frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}_i$, provide initial values for β. Note that in practice $\underline{\hat{\beta}}$ and $\hat{\alpha}_i$ may be computed as $\underline{\hat{\beta}} = (\sum_{i=1}^{n} \underline{X}_i' M_{T_i} \underline{X}_i)^{-1} \sum_{i=1}^{n} \underline{X}_i' M_{T_i} Y_i$ (within OLS estimator) and $\hat{\alpha}_i = \frac{1}{T_i} e_{T_i}' (Y_i - \underline{X}_i \underline{\hat{\beta}})$, where $M_{T_i} = I_{T_i} - \frac{1}{T_i} J_{T_i}$, i.e. the within transformation matrix. See Balestra (1996) for details.
2. Run the OLS regression $\phi_\nu^{-1}(\underline{\hat{u}}_{it}^2) = Z_{it}^1 \gamma_1 +$ residuals $(i = 1, 2, \ldots, n;$ $t = 1, 2, \ldots, T_i)$, where $\underline{\hat{u}}_{it} = Y_{it} - \hat{\alpha}_i - \underline{X}_{it} \underline{\hat{\beta}}$ and $\phi_\nu^{-1}(\cdot)$ is the (supposed well-defined) inverse function of $\phi_\nu(\cdot)$. The non-intercept

parameters of $\hat{\gamma}_1$ and the intercept parameter of $\hat{\gamma}_1$ minus γ_{1_c}, where γ_{1_c} is an intercept correction term, give initial values for γ_1. The desirability of an intercept correction of $\hat{\gamma}_1$ arises from the fact that, even if we suppose that $\underline{\hat{u}}_{it}$ is equal to the true disturbance ν_{it}, the (conditional) expectation of the error term in the above OLS regression is usually not zero (and even not necessarily a constant). The "optimal" value of the intercept correction term γ_{1_c} depends upon the functional form $\phi_\nu^{-1}(\cdot)$ and the actual distribution of the ν_{it}. In the case of the multiplicative heteroscedasticity formulation where $\phi_\nu^{-1}(\cdot)$ is simply equal to $\ln(\cdot)$, a sensible choice is $\gamma_{1_c} = -1.2704$. This follows from the fact that $E[\ln(\nu_{it}^2) - \ln(\sigma_{\nu_{it}}^2)] = E[\ln(\nu_{it}^2/\sigma_{\nu_{it}}^2)] = -1.2704$ if $\nu_{it} \sim N(0, \sigma_{\nu_{it}}^2)$; see Harvey (1976).

3. Finally, run the OLS regression $\phi_\mu^{-1}((\hat{\alpha}_i - \bar{\alpha})^2) = Z_i^2\gamma_2 +$ residuals $(i = 1, 2, \ldots, n)$, where $\phi_\mu^{-1}(\cdot)$ is the (supposed well-defined) inverse function of $\phi_\mu(\cdot)$. According to the same reasoning as above, the non-intercept parameters of $\hat{\gamma}_2$ and the intercept parameter of $\hat{\gamma}_2$ minus γ_{2_c}, where γ_{2_c} is an intercept correction term, give initial values for γ_2. In the case of the multiplicative heteroscedasticity formulation where $\phi_\mu^{-1}(\cdot)$ is again equal to $\ln(\cdot)$, γ_{2_c} should also be set to -1.2704.

Note that a simpler alternative to the step 2 and 3 is workable. It merely consists in computing the "mean variance components" $\hat{\sigma}_\nu^2 = \frac{1}{N}\sum_{i=1}^n \sum_{t=1}^{T_i} \underline{\hat{u}}_{it}^2$ and $\hat{\sigma}_\mu^2 = \frac{1}{n}\sum_{i=1}^n (\hat{\alpha}_i - \bar{\alpha})^2$. The inverse function values $\phi_\nu^{-1}(\hat{\sigma}_\nu^2)$ and $\phi_\mu^{-1}(\hat{\sigma}_\mu^2)$ may then be used for the first elements (intercepts) of γ_1 and γ_2, their remaining elements being simply set equal to zero.

References

Balestra, P. (1996), "Fixed effect models and fixed coefficient models", in: Mátyás, L., Sevestre, P., editors, *The Econometrics of Panel Data*, Kluwer Academic Publishers, Dordrecht. Ch. 3.

Baltagi, B.H. (1988), "An alternative heteroscedastic error components model", Problem 88.2.2, *Econometric Theory*, Vol. 4, pp. 349–350.

Baltagi, B.H. (1995), *Econometric Analysis of Panel Data*, John Wiley & Sons, New York.

Baltagi, B.H., Griffin, J.M. (1988), "A generalized error component model with heteroscedastic disturbances", *International Economic Review*, Vol. 29 (4), pp. 745–753.

Baltagi, B.H., Bresson, G., Pirotte, A. (2004), "Joint LM test for homoscedasticity in a one-way error component model", *Journal of Economics*. in preparation.

Bollerslev, T., Wooldridge, J. (1992), "Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances", *Econometric Reviews*, Vol. 11 (2), pp. 143–172.

Breusch, T.S. (1987), "Maximum likelihood estimation of random effects models", *Journal of Econometrics*, Vol. 36, pp. 383–389.

Breusch, T.S., Pagan, A.R. (1979), "A simple test for heteroscedasticity and random coefficient variation", *Econometrica,* Vol. 47 (5), pp. 1287–1297.

Chamberlain, G. (1987), "Asymptotic efficiency in estimation with conditional moment restrictions", *Journal of Econometrics,* Vol. 34, pp. 305–334.

Cox, D.R. (1961), "Tests of separate families of hypotheses", in: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, University of California Press, Berkeley, pp. 105–123.

Cox, D.R. (1962), "Further results on tests of separate families of hypotheses", *Journal of the Royal Statistical Society, Series B,* Vol. 24, pp. 406–424.

Davidson, R., MacKinnon, J.G. (1981), "Several tests for model specification in the presence of alternative hypotheses", *Econometrica,* Vol. 49 (3), pp. 781–793.

Gourieroux, C., Monfort, A. (1993), "Pseudo-likelihood methods", in: Maddala, G.S., Rao, C.R., Vinod, H.D., editors, *Handbook of Statistics, Vol. 11*, North-Holland, Amsterdam. Ch. 12.

Gourieroux, C., Monfort, A., Trognon, A. (1984), "Pseudo-maximum likelihood methods: theory", *Econometrica,* Vol. 52 (3), pp. 681–700.

Harvey, A.C. (1976), "Estimating regression models with multiplicative heteroscedasticity", *Econometrica,* Vol. 44 (3), pp. 461–465.

Holly, A., Gardiol, L. (2000), "A score test for individual heteroscedasticity in a one-way error components model", in: Kirshnakumar, J., Ronchetti, E., editors, *Panel Data Econometrics: Future Directions*, Elsevier, Amsterdam. Ch. 10.

Lejeune, B. (1998), "Error components models and variable heterogeneity: modelisation, second order pseudo-maximum likelihood estimation and specification testing", Ph.D. dissertation, Université de Liège and Université de Paris XII-Val de Marne.

Lejeune B. (2004), "A distribution-free joint test and one-directional robust tests for random individual effects and heteroscedasticity allowing for unbalanced panels", Working Paper, mimeo, University of Liège, Liège.

Li, Q., Stengos, T. (1994), "Adaptive estimation on the panel data error component model with heteroscedasticity of unknown form", *International Economic Review,* Vol. 35, pp. 981–1000.

Magnus, J.R. (1978), "Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix", *Journal of Econometrics,* Vol. 7, pp. 281–312. Corrigenda, *Journal of Econometrics,* Vol. 10 (1978) p. 261.

Magnus, J.R. (1982), "Multivariate error components analysis of linear and non-linear regression models by maximum likelihood", *Journal of Econometrics,* Vol. 19, pp. 239–285.

Magnus, J.R., Neudecker, H. (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, New York.

Mazodier, P., Trognon, A. (1978), "Heteroscedasticity and stratification in error components models", *Annales de l'INSEE,* Vol. 30–31, pp. 451–482.

Newey, W. (1985), "Maximum likelihood specification testing and conditional moment tests", *Econometrica,* Vol. 53 (5), pp. 1047–1070.

Newey, W. (1990), "Semiparametric efficiency bounds", *Journal of Applied Econometrics,* Vol. 5, pp. 99–135.

Newey, W. (1993), "Efficient estimation of models with conditional moment restrictions", in: Maddala, G.S., Rao, C.R., Vinod, H.D., editors, *Handbook of Statistics, Vol. 11*, North-Holland, Amsterdam. Ch. 16.

Phillips, R.L. (2003), "Estimation of a stratified error components model", *International Economic Review,* Vol. 44, pp. 501–521.

Quandt, R.E. (1983), “Computational problems and methods”, in: Griliches, Z., Intriligator, M.D., editors, *Handbook of Econometrics, Vol. 1*, North-Holland, Amsterdam. Ch. 12.

Randolph, W.C. (1988), “A transformation for heteroscedastic error components regression models”, *Economics Letters,* Vol. 27, pp. 349–354.

Rao, S.R.S., Kaplan, J., Cochran, W.G. (1981), “Estimators for the one-way random effects model with unequal error variances”, *Journal of the American Statistical Association*, Vol. 76, pp. 89–97.

Roy, N. (2002), “Is adaptive estimation useful for panel models with heteroscedasticity in the individual specific error component? Some Monte Carlo evidence”, *Econometric Reviews,* Vol. 21, pp. 189–203.

Tauchen, G. (1985), “Diagnostic testing and evaluation of maximum likelihood models”, *Journal of Econometrics,* Vol. 30, pp. 415–443.

Verbon, H.A.A. (1980), “Testing for heteroscedasticity in a model of seemingly unrelated regression equations with variance components”, *Economics Letters,* Vol. 5, pp. 149–153.

Wansbeek, T. (1989), “An alternative heteroscedastic error components model”, Solution 88.1.1, *Econometric Theory,* Vol. 5, p. 326.

White, H. (1987), “Specification testing in dynamic models”, in: Bewley, T., editor, *Advances in Econometrics – Fifth World Congress, Vol. 1*, Cambridge University Press, New York. Ch. 1.

White, H. (1994), *Estimation, Inference and Specification Analysis,* Cambridge University Press, Cambridge.

Wooldridge, J. (1990), “A unified approach to robust, regression-based specification tests”, *Econometric Theory,* Vol. 6, pp. 17–43.

Wooldridge, J. (1991a), “On the application of robust, regression-based diagnostics to models of conditional means and conditional variances”, *Journal of Econometrics,* Vol. 47, pp. 5–46.

Wooldridge, J. (1991b), “Specification testing and quasi-maximum likelihood estimation”, *Journal of Econometrics,* Vol. 48, pp. 29–55.

Wooldridge, J. (1994), “Estimation and inference for dependent processes”, in: Engle, R., McFadden, D., editors, *Handbook of Econometrics, Vol. 4*, North-Holland, Amsterdam. Ch. 45.

Wooldridge, J. (1995), “Selection corrections for panel data models under conditional mean independence assumptions”, *Journal of Econometrics,* Vol. 68, pp. 115–132.

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74003-2

CHAPTER 3

Finite Sample Properties of FGLS Estimator for Random-Effects Model under Non-Normality

Aman Ullah* and Xiao Huang

Department of Economics, University of California, Riverside, CA 92521-0427, USA
E-mail address: aman.ullah@ucr.edu; xiao.huang@email.ucr.edu

Abstract

This paper considers the finite sample properties of the feasible generalized least square (FGLS) estimator for the random-effects model with non-normal errors. By using the asymptotic expansion, we study the effects of skewness and excess kurtosis on the bias and Mean Square Error (MSE) of the estimator. The numerical evaluation of our results is also presented.

Keywords: finite sample, non-normality, panel data, random-effects

JEL classifications: C1, C4

3.1 Introduction

In random-effects (error component) models when variances of the individual-specific effect and error term are unknown, feasible generalized least square (FGLS) is the standard way for estimation (Baltagi, 2001). For large sample size, FGLS has the same asymptotic efficiency as the GLS estimator when variances are known (Fuller and Battese, 1974). However, we deal with data sets of small and moderately large sample size in many situations and the disturbances are typically believed to be non-normally distributed.

Maddala and Mount (1973) provided a simulation study on the efficiency of slope estimators for a static one-way error component panel data model. They considered both normal and non-normal errors in simulations, where their non-normal errors are from lognormal distribution.

* Corresponding author.

It is found that maximum likelihood estimator performs as well as other types of FGLS estimators under both normal and lognormal errors in small samples and all estimators give equally well results. Baltagi (1981) investigated thoroughly various estimation and testing procedures in a static two-way error component model and extended many estimation results in one-way models to two-way models. Taylor (1980) examined the exact analytical small sample efficiency of FGLS estimator compared to between groups estimator and within groups estimator under the assumption of normality.

Despite of previous studies, there has been no analytical result on how non-normality affects the statistical properties of FGLS estimator in static panel data model when sample size is finite. Further, we note that the exact analytical result for the non-normal case is difficult to obtain and it needs the specification of the form of the non-normal distribution. This paper gives the large-n (fixed T) approximate analytical result of finite sample behavior of FGLS with non-normal disturbances. We derive the approximate bias, up to $\mathrm{O}(1/n)$, and the mean square error (MSE), up to $\mathrm{O}(1/n^2)$, of the FGLS estimator in a static regression model under the assumption that the first four moments of the errors are finite. For the case of dynamic panel, the finite sample properties has been studied in several papers through simulation, for example, Nerlove (1967, 1971), Arellano and Bond (1991), and Kiviet (1995), and they are not directly related to the static case studied in this paper.

The paper is organized as follows. Section 3.2 gives the main results. In Section 3.3 are detailed proofs. Some numerical results are given and discussed in Section 3.4. Section 3.5 provides the conclusion.

3.2 Main results

Let us consider the following random effect model,

$$\begin{aligned} y_{it} &= x_{it}\beta + w_{it}, \\ w_{it} &= \alpha_i + u_{it}, \qquad i = 1, \ldots, n, t = 1, \ldots, T, \end{aligned} \tag{3.1}$$

where y_{it} is the dependent variable, x_{it} is an $1 \times k$ vector of exogenous variables, β is a $k \times 1$ coefficient vector and the error w_{it} consists of a time-invariant random component, α_i, and a random component u_{it}. We can also write the above model in a vector form as

$$\begin{aligned} y &= X\beta + w, \\ w &= D\alpha + u, \\ D &= I_n \otimes \iota_T, \end{aligned} \tag{3.2}$$

where y is $nT \times 1$, X is $nT \times k$, w is $nT \times 1$, α is $n \times 1$, I_n is an identity matrix of dimension n, and ι_T is $T \times 1$ with all elements equal to one.

We assume both α_i and u_{it} are i.i.d. and mutually independent and

$$\begin{aligned}
E\alpha_i &= 0, \qquad E\alpha_i^2 = \sigma_\alpha^2,\\
E\alpha_i^3 &= \sigma_\alpha^3\gamma_{1\alpha}, \qquad E\alpha_i^4 = \sigma_\alpha^4(\gamma_{2\alpha}+3),\\
E\alpha_i\alpha_j &= \begin{cases}\sigma_\alpha^2 & \text{if } i=j,\\ 0 & \text{if } i\neq j,\end{cases}\\
Eu_{it} &= 0, \qquad Eu_{it}^2 = \sigma_u^2,\\
Eu_{it}^3 &= \sigma_u^3\gamma_{1u}, \qquad Eu_{it}^4 = \sigma_u^4(\gamma_{2u}+3),\\
Eu_{it}u_{js} &= \begin{cases}\sigma_u^2 & \text{if } i=j, t=s,\\ 0 & \text{otherwise,}\end{cases}\\
E\alpha_i x_{it} &= Eu_{js}x_{it} = 0, \quad i,j = 1,\ldots,n \text{ and } s,t = 1,\ldots,T,
\end{aligned} \tag{3.3}$$

where $\gamma_{1\alpha}, \gamma_{1u}$ and $\gamma_{2\alpha}, \gamma_{2u}$ are Pearson's measures of skewness and kurtosis of the distribution.

The variance–covariance matrix of w can be written as

$$\begin{aligned}
Eww' &= \sigma_u^2\big(Q + \lambda^{-1}\overline{Q}\big)\\
&= \sigma_u^2\Omega^{-1},
\end{aligned} \tag{3.4}$$

where $Q = I_{nT} - \overline{Q}$, $\overline{Q} = DD'/T$, $\lambda = \sigma_u^2/\sigma_\eta^2$ and $\sigma_\eta^2 = \sigma_u^2 + T\sigma_\alpha^2$, $0 < \lambda \leq 1$. Obviously, we have the following properties of Q and $\overline{Q}$:

$$\begin{aligned}
&Q^2 = Q, \qquad \overline{Q}^2 = \overline{Q}, \qquad Q\overline{Q} = 0, \quad \text{and}\\
&\Omega = Q + \lambda\overline{Q} = I_{nT} - (1-\lambda)\overline{Q}.
\end{aligned} \tag{3.5}$$

The generalized least square (GLS) estimator of β when the variances of u_{it} and α_i are known is given by

$$\hat{\beta}_{\text{GLS}} = (X'\Omega X)^{-1}X'\Omega y. \tag{3.6}$$

When the variances of u_{it} and α_i are unknown, then feasible GLS estimator is used by replacing Ω with its estimator, $\widehat{\Omega}$,

$$\hat{\beta}_{\text{FGLS}} = (X'\widehat{\Omega}X)^{-1}X'\widehat{\Omega}y, \tag{3.7}$$

where

$$\widehat{\Omega} = Q + \hat{\lambda}\overline{Q}, \tag{3.8}$$

$$\hat{\sigma}_u^2 = \frac{u'(Q - QX(X'QX)^{-1}X'Q)u}{n(T-1)-k}, \tag{3.9}$$

$$\hat{\sigma}_\eta^2 = \frac{w'(\overline{Q} - \overline{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q})w}{n-k}, \tag{3.10}$$

$$\hat{\lambda} = \hat{\sigma}_u^2/\hat{\sigma}_\eta^2. \tag{3.11}$$

By expanding the terms in (3.8)–(3.11) and plugging them into (3.7), we obtain the analytical expression of the second-order bias and mean square error for $\hat{\beta}_{\mathrm{FGLS}}$. The detailed proofs are given in Section 3.3 and we give the main result in the following theorem.

THEOREM 3.1. *Under assumption (3.3) the large-sample asymptotic approximations for the bias vector $E(\hat{\beta}_{\mathrm{FGLS}} - \beta)$ up to $\mathrm{O}(n^{-1})$ and mean square error matrix $E((\hat{\beta}_{\mathrm{FGLS}} - \beta)(\hat{\beta}_{\mathrm{FGLS}} - \beta)')$ up to $\mathrm{O}(n^{-2})$ are given by*

$$\mathrm{Bias} = \frac{\lambda(1-\lambda)}{n^2}\left(\frac{\sigma_u\gamma_{1u}}{T} - \sigma_\alpha\gamma_{1\alpha}\right)\left(A^{-1} - \lambda A^{-1}BA^{-1}\right)X'\iota_{nT},$$

$$\mathrm{MSE} = \sigma_u^2(X'\Omega X)^{-1} + \frac{\lambda\sigma_u^2}{n^2}\left[\frac{2T}{T-1} - \frac{\gamma_{2u}}{T}(1-\lambda)^2 - \gamma_{2\alpha}(1-\lambda)^2\right]\Delta + \frac{C}{n} + \frac{C'}{n},$$

where ι_{nT} is an $nT \times 1$ vector of ones, $A = \frac{1}{n}X'\Omega X$, $B = \frac{1}{n}X'\overline{Q}X$, $\Delta = A^{-1}(B - \lambda BA^{-1}B)A^{-1}$, and

$$C = \frac{\lambda\sigma_u^2}{n}A^{-1}\frac{X'\Omega}{\sqrt{n}}\Bigg[\lambda\gamma_{2u}\left(I \odot \overline{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q}\right) - \frac{\gamma_{2u}}{T-1}\left(I \odot QX(X'QX)^{-1}X'Q\right) + \frac{\gamma_{2\alpha}(1-\lambda)^2}{\lambda T^2}D\left(I \odot DX(X'\overline{Q}X)^{-1}X'D\right)D'\Bigg]P_1'A^{-1},$$

in which $P_1 = (X'\overline{Q} - BA^{-1}X'\Omega)/\sqrt{n}$.

The proof of Theorem 3.1 is given in Section 3.3. When errors are normally distributed, $\gamma_{1\alpha} = \gamma_{2\alpha} = \gamma_{1u} = \gamma_{2u} = 0$ and we get

COROLLARY 3.1. *Under assumption (3.3), when the errors are normally distributed, the large-sample asymptotic approximations for the bias vector $E(\hat{\beta}_{\mathrm{FGLS}} - \beta)$ up to $\mathrm{O}(n^{-1})$ and mean square error matrix $E((\hat{\beta}_{\mathrm{FGLS}} - \beta)(\hat{\beta}_{\mathrm{FGLS}} - \beta)')$ up to $\mathrm{O}(n^{-2})$ are given by*

$$\mathrm{Bias} = 0,$$

$$\mathrm{MSE} = \sigma_u^2(X'\Omega X)^{-1} + \frac{2\lambda\sigma_u^2 T}{n^2(T-1)}\Delta.$$

If the non-normality comes from α_i, not from u_{it}, then $\gamma_{1u} = \gamma_{2u} = 0$ and we have

COROLLARY 3.2. *Under assumption* (3.3), *when only α_i is non-normally distributed, the large-sample asymptotic approximations for the bias vector $E(\hat{\beta}_{\text{FGLS}} - \beta)$ up to $\text{O}(n^{-1})$ and mean square error matrix $E((\hat{\beta}_{\text{FGLS}} - \beta)(\hat{\beta}_{\text{FGLS}} - \beta)')$ up to $\text{O}(n^{-2})$ are given by*

$$\text{Bias} = -\frac{\lambda(1-\lambda)\sigma_\alpha\gamma_{1\alpha}}{n^2}\big(A^{-1} - \lambda A^{-1}BA^{-1}\big)X'\iota_{nT},$$

$$\text{MSE} = \sigma_u^2(X'\varOmega X)^{-1} + \frac{\lambda\sigma_u^2}{n^2}\left[\frac{2T}{T-1} - \gamma_{2\alpha}(1-\lambda)^2\right]\varDelta + \frac{F}{n} + \frac{F'}{n},$$

where

$$F = \frac{\lambda\sigma_u^2}{n}A^{-1} \times \frac{X'\varOmega}{\sqrt{n}}\left[\frac{\gamma_{2\alpha}(1-\lambda)^2}{\lambda T^2}D\big(I \odot DX(X'\overline{Q}X)^{-1}X'D\big)D'\right]P_1'A^{-1}.$$

Similarly, if the non-normality comes only from u_{it}, then $\gamma_{1\alpha} = \gamma_{2\alpha} = 0$ and we have

COROLLARY 3.3. *Under assumption* (3.3), *when only u_{it} is non-normally distributed, the large-sample asymptotic approximations for the bias vector $E(\hat{\beta}_{\text{FGLS}} - \beta)$ up to $\text{O}(n^{-1})$ and mean square error matrix $E((\hat{\beta}_{\text{FGLS}} - \beta)(\hat{\beta}_{\text{FGLS}} - \beta)')$ up to $\text{O}(n^{-2})$ are given by*

$$\text{Bias} = \frac{\lambda(1-\lambda)\sigma_u\gamma_{1u}}{n^2T}\big(A^{-1} - \lambda A^{-1}BA^{-1}\big)X'\iota_{nT},$$

$$\text{MSE} = \sigma_u^2(X'\varOmega X)^{-1} + \frac{\lambda\sigma_u^2}{n^2}\left[\frac{2T}{T-1} - \frac{\gamma_{2u}}{T}(1-\lambda)^2\right]\varDelta + \frac{G}{n} + \frac{G'}{n},$$

where

$$G = \frac{\lambda\sigma_u^2}{n}A^{-1}\frac{X'\varOmega}{\sqrt{n}}\left[\lambda\gamma_{2u}\big(I \odot \overline{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q}\big) - \frac{\gamma_{2u}}{T-1}\big(I \odot QX(X'QX)^{-1}X'Q\big)\right]P_1'A^{-1}.$$

We note that the asymptotic MSE of $\hat{\beta}_{\text{FGLS}}$ is given by $\sigma_u^2(X'\varOmega X)^{-1}$. The following remarks follow from the results in Theorem 3.1 and Corollary 3.1.

REMARK 3.1. The Bias depends only on skewness coefficient. Bias is zero if $\lambda = 1$ or $\lambda = 0$, where $\lambda = 1$ implies $\sigma_\alpha^2 = 0$ and $\lambda = 0$ implies $\sigma_u^2 = 0$. Also note that for symmetric distributions, $\gamma_{1\alpha} = \gamma_{1u} = 0$, or for distributions satisfying $\gamma_{1u}/\gamma_{1\alpha} = T\sigma_\alpha/\sigma_u$, Bias is zero. Consider the

Table 3.1. ***$n = 10$, $T = 5$, α is non-normal and u is normal, $\sigma_\alpha = 1$, $\sigma_u = 0.6$***

θ_2	$\gamma_{1\alpha}$	$\gamma_{2\alpha}$	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	0.013656	0	0.002601	0.002902
2.0	−1.70	5.64	0.008769	0	0.002848	0.002902
2.5	−1.23	2.93	0.006350	0	0.002921	0.002902
3.0	−0.94	1.74	0.004853	0	0.002954	0.002902

Table 3.2. ***$n = 50$, $T = 5$, α is non-normal and u is normal, $\sigma_\alpha = 1$, $\sigma_u = 0.6$***

θ_2	$\gamma_{1\alpha}$	$\gamma_{2\alpha}$	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	0.002661	0	0.000580	0.000594
2.0	−1.70	5.64	0.001709	0	0.000591	0.000594
2.5	−1.23	2.93	0.001237	0	0.000594	0.000594
3.0	−0.94	1.74	0.000946	0	0.000596	0.000594

term

$$\begin{aligned} A^{-1} - \lambda A^{-1}BA^{-1} &= A^{-1}(I_{nT} - \lambda B)A^{-1} \\ &= \left(\frac{X'\Omega X}{n}\right)^{-1}(X'QX)\left(\frac{X'\Omega X}{n}\right)^{-1} \geq 0. \end{aligned}$$

Thus $A^{-1} - \lambda A^{-1}BA^{-1}$ is a positive semidefinite matrix. Therefore, provided $X'\iota_{nT} \geq 0$,

$$\text{Bias} \gtrless 0 \quad \text{if } \frac{\gamma_{1u}}{\gamma_{1\alpha}} \gtrless \frac{T\sigma_\alpha}{\sigma_u},$$

$$\frac{\partial \text{Bias}}{\partial \gamma_{1u}} \geq 0, \qquad \frac{\partial \text{Bias}}{\partial \gamma_{1\alpha}} \leq 0, \quad \text{and} \quad \frac{\partial^2 \text{Bias}}{\partial \gamma_{1u}\partial \gamma_{1\alpha}} = 0.$$

For the nature of decreasing slope of bias with respect to $\gamma_{1\alpha}$, see Tables 3.1 to 3.3. This Bias direction does not hold, that is bias direction is not determined, if each element of $X'\iota_{nT}$ is not positive or negative.

REMARK 3.2. Under certain restrictions, there are also some monotonic relations between the Bias and the variances of the error components. Consider the Bias expression in Corollary 3.2, where only α is non-normally distributed. For simplicity, let $k = 1$ and $H = (X'QX)(X'\Omega X/n)^{-1}$. The derivative of the Bias w.r.t. σ_α^2 gives

$$\frac{\partial \text{Bias}}{\partial \sigma_\alpha^2} = -\frac{\gamma_{1\alpha}}{n^2}\left[H\frac{\partial \lambda(1-\lambda)\sigma_\alpha}{\partial \sigma_\alpha^2} + \lambda(1-\lambda)\sigma_a \frac{\partial H}{\partial \sigma_\alpha^2}\right]X'\iota,$$

where

$$\partial\lambda(1-\lambda)\sigma_\alpha/\partial\sigma_\alpha^2 = 2\sigma_\alpha^{-1}\lambda(1-\lambda)(4\lambda-1)\begin{matrix}\geq 0 & \text{if } \lambda \geq 1/4, \\ < 0 & \text{if } \lambda < 1/4,\end{matrix}$$

$$\partial H/\partial\sigma_\alpha^2 = 2n^{-1}T\sigma_u^{-2}\lambda^2(X'QX)(X'\Omega X/n)^{-3}(X'\overline{Q}X) \geq 0.$$

For $X'\iota > 0$, if $\gamma_{1\alpha} < 0$, Bias is an increasing function of σ_α^2 when $\lambda \geq 1/4$. When $\lambda < 1/4$, the monotonicity is not determined.

Similar result holds for $\partial\text{Bias}/\partial\sigma_u^2$. For $X'\iota > 0$, if $\gamma_{1u} < 0$, it is found that Bias is an increasing function of σ_u^2 when $\lambda > 3/4$. When $\lambda \leq 3/4$, the monotonicity is again not determined.

REMARK 3.3. Under the non-normality of errors, the MSE depends only on kurtosis. The approximate MSE for normal distribution is greater than or equal to asymptotic MSE, i.e.

$$\sigma_u^2(X'QX)^{-1} + \frac{2\lambda\sigma_u^2 T}{n^2(T-1)} \geq \sigma_u^2(X'QX)^{-1}.$$

The results for approximate MSE result under both normal and non-normal errors in Tables 3.1 to 3.7 suggest that the asymptotic MSE results are generally the same as the approximate MSE results for moderately large samples, at least up to 4 digits.

3.3 Derivation

PROOF OF THEOREM 3.1. The expansion of the bias vector follows directly from the expansion of $\hat{\beta}_{\text{FGLS}}$ around its true value, β. From (3.7) we know that the expansion of $\hat{\beta}_{\text{FGLS}}$ requires the expansion of $\hat{\lambda}$, which further involves the expansion of $\hat{\sigma}_u^2$ and $\hat{\sigma}_\eta^2$. Let us start with the Taylor series expansion of $\hat{\sigma}_u^2$ and $\hat{\sigma}_\eta^2$. From (3.9), we have

$$\begin{aligned}\hat{\sigma}_u^2 &= \frac{u'Qu - u'QX(X'QX)^{-1}X'Qu}{n(T-1)-k} \\ &= \frac{1}{n(T-1)}\left[1 - \frac{k}{n(T-1)}\right]^{-1}\left[\sigma_u^2 n(T-1)\left(1 + \frac{v_u}{\sqrt{n}}\right) - \sigma_u^2 v_u^*\right]\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n(T-1)}\left[1+\frac{k}{n(T-1)}+\frac{k^2}{n^2(T-1)^2}+\cdots\right]\\
&\quad \times\left[\sigma_u^2 n(T-1)\left(1+\frac{v_u}{\sqrt{n}}\right)-\sigma_u^2 v_u^*\right]\\
&= \sigma_u^2\left[1+\frac{v_u}{\sqrt{n}}+\frac{k-v_u^*}{n(T-1)}\right]+\mathrm{O}_p\left(n^{-3/2}\right),
\end{aligned} \tag{3.12}$$

where

$$v_u = \sqrt{n}\left(\frac{u'Qu}{n(T-1)\sigma_u^2}-1\right), \tag{3.13}$$

$$v_u^* = u'QX(X'QX)^{-1}XQu/\sigma_u^2. \tag{3.14}$$

Both v_u and v_u^* are $\mathrm{O}_p(1)$. Similarly, we define other $\mathrm{O}_p(1)$ terms frequently used in the proof,

$$v_\alpha = \sqrt{n}\left(\alpha'\alpha\sigma_\alpha^{-2}n^{-1}-1\right), \tag{3.15}$$

$$\varepsilon_u = \sqrt{n}\left(u'\overline{Q}u\sigma_u^{-2}n^{-1}-1\right), \tag{3.16}$$

$$v_\alpha^* = \alpha'D'X(X'\overline{Q}X)^{-1}X'D\alpha/\sigma_\eta^2, \tag{3.17}$$

$$\varepsilon_u^* = u'\overline{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q}u/\sigma_\eta^2, \tag{3.18}$$

$$v_{\alpha u} = \frac{u'D\alpha}{\sqrt{n}\sigma_\eta^2}, \tag{3.19}$$

$$v_{\alpha u}^* = \alpha'D'X(X'\overline{Q}X)^{-1}X'\overline{Q}u/\sigma_\eta^2. \tag{3.20}$$

For $\hat{\sigma}_\eta^2$, we have

$$\begin{aligned}
w'\overline{Q}w &= \alpha'D'\overline{Q}D\alpha + u'\overline{Q}u + 2u'\overline{Q}D\alpha\\
&= nT\sigma_\alpha^2\left(1+v_\alpha/\sqrt{n}\right)+n\sigma_u^2\left(1+\varepsilon_u/\sqrt{n}\right)+2\sqrt{n}\sigma_\eta^2 v_{\alpha u}\\
&= \sigma_\eta^2\left[n+\sqrt{n}\left((1-\lambda)v_\alpha+\lambda\varepsilon_u+2v_{\alpha u}\right)\right],
\end{aligned} \tag{3.21}$$

$$w'\overline{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q}w = \sigma_\eta^2\left(v_\alpha^*+\varepsilon_u^*+2v_{\alpha u}^*\right). \tag{3.22}$$

Now plug (3.21) and (3.22) into (3.10) along with $1/(n-k) = 1/n + k/n^2 + \cdots$, we get

$$\hat{\sigma}_\eta^2 = \sigma_\eta^2\left[1+\frac{(1-\lambda)v_\alpha+\lambda\varepsilon_u+2v_{\alpha u}}{\sqrt{n}}+\frac{k-(v_\alpha^*+\varepsilon_u^*+2v_{\alpha u}^*)}{n}\right] + \mathrm{O}_p\left(n^{-3/2}\right). \tag{3.23}$$

Using (3.12) and (3.23), it can be verified that

$$\hat{\lambda} = \lambda\left[1+\frac{f}{\sqrt{n}}+\frac{f^*-fv_u+f^2}{n}\right]+\mathrm{O}_p\left(n^{-3/2}\right), \tag{3.24}$$

where

$$\begin{aligned} f &= v_u - (1-\lambda)v_\alpha - \lambda\varepsilon_u - 2v_{\alpha u} \\ &= \frac{1}{\sqrt{n}\sigma_u^2}u'\left(\frac{Q}{T-1} - \lambda\overline{Q}\right)u - (1-\lambda)\left(\frac{\alpha'\alpha}{\sqrt{n}\sigma_\alpha^2}\right) - \frac{2}{\sqrt{n}\sigma_\eta^2}u'D\alpha, \end{aligned} \tag{3.25}$$

$$f^* = v_\alpha^* + \varepsilon_u^* + 2v_{\alpha u}^* - v_u^*/(T-1) - k(T-2)/(T-1). \tag{3.26}$$

Multiplying both sides of (3.24) by $\sqrt{n}$ and rearranging the equation gives

$$\sqrt{n}(\hat{\lambda} - \lambda) = \lambda f + \lambda\left(f^* - fv_u + f^2\right)/\sqrt{n} + \mathrm{O}_p\left(n^{-1}\right). \tag{3.27}$$

Now define

$$\delta = \sqrt{n}(\hat{\lambda} - \lambda) \tag{3.28}$$

so that $\delta^2 = \lambda^2 f^2 + \mathrm{O}_p(n^{-1/2})$. Using the above definition, we have

$$\widehat{\Omega} = \Omega + \overline{Q}\delta/\sqrt{n}, \tag{3.29}$$

$$X'\widehat{\Omega}X/n = A + B\delta/\sqrt{n}. \tag{3.30}$$

Now plug (3.29) and (3.30) into (3.7) and multiply both sides by $\sqrt{n}$ we have

$$\begin{aligned} &\sqrt{n}(\hat{\beta}_{\text{FGLS}} - \beta) \\ &\quad = \left(A + B\delta/\sqrt{n}\right)^{-1}\left[X'\left(\Omega + \overline{Q}\delta/\sqrt{n}\right)w/\sqrt{n}\right] \\ &\quad = A^{-1}\left(X'\Omega w/\sqrt{n}\right) + A^{-1}P_1 w\delta/\sqrt{n} + A^{-1}P_2 w\delta^2/n, \end{aligned} \tag{3.31}$$

where P_1 is as given in Theorem 3.1 and $P_2 = -BA^{-1}P_1$. It can be easily verified that $\overline{Q}\iota_{nT} = \iota_{nT}$, $P_1 X = P_2 X = 0$, $P_1\iota_{nT} = (X' - \lambda BA^{-1}X')\iota_{nT}/\sqrt{n}$, $P_1\overline{Q}X/\sqrt{n} = B - \lambda BA^{-1}B$, and $P_1 QX/\sqrt{n} = -P_1\overline{Q}X/\sqrt{n}$.

Then using (3.28) we get

$$\sqrt{n}(\hat{\beta}_{\text{FGLS}} - \beta) = \xi_0 + \xi_{-1/2} + \xi_{-1} + \mathrm{O}_p\left(n^{-3/2}\right), \tag{3.32}$$

where

$$\begin{aligned} \xi_0 &= A^{-1}\left(X'\Omega w/\sqrt{n}\right), \qquad \xi_{-1/2} = \lambda A^{-1}P_1 wf/\sqrt{n}, \\ \xi_{-1} &= \lambda^2 A^{-1}P_2 wf^2/n + \lambda A^{-1}P_1 w\left(f^* - fv_u + f^2\right)/n. \end{aligned}$$

Taking expectation of (3.32) to obtain the bias vector up to $\mathrm{O}_p(n^{-1/2})$

$$\begin{aligned} E\left[\sqrt{n}(\hat{\beta}_{\text{FGLS}} - \beta)\right] &= E\xi_0 + E\xi_{-1/2} \\ &= \lambda A^{-1}P_1 E(wf)/\sqrt{n}. \end{aligned} \tag{3.33}$$

It is easy to verify that $E\xi_0 = 0$. Now let us evaluate $E(wf) = E(D\alpha + u)f = DE(\alpha f) + E(uf)$. From (3.25) we get

$$\begin{aligned} DE(\alpha f) &= DE\big[\sigma_u^{-2} u'\big(Q/(T-1) - \lambda \widehat{Q}\big) u\alpha/\sqrt{n} \\ &\qquad -(1-\lambda)\sigma_\alpha^{-2}\alpha'\alpha\alpha/\sqrt{n} - 2\sigma_\eta^{-2} u' D\alpha\alpha/\sqrt{n}\big] \\ &= -(1-\lambda)\sigma_\alpha^{-2} E(\alpha'\alpha \cdot \alpha)/\sqrt{n} \\ &= -(1-\lambda)\gamma_{1\alpha}\sigma_\alpha \iota_{nT}/\sqrt{n}, \end{aligned} \tag{3.34}$$

$$\begin{aligned} E(uf) &= E\big[u(v_u - \lambda\varepsilon_u)\big] \\ &= \sigma_u^{-2}\left[E(u'u \cdot u)/(T-1) - \frac{\lambda(T-1)+1}{T-1} E(u'\bar{Q}u \cdot u)\right]\Big/\sqrt{n} \\ &= \gamma_{1u}\sigma_u(1-\lambda)T^{-1}n^{-1/2}\iota_{nT}. \end{aligned} \tag{3.35}$$

Combine (3.34) and (3.35) we have

$$E(wf) = (1-\lambda)(\gamma_{1u}\sigma_u/T - \gamma_{1\alpha}\sigma_\alpha)\iota_{nT}/\sqrt{n}. \tag{3.36}$$

Hence substituting (3.36) in (3.33) we get the bias result in Theorem 3.1.

The mean square error matrix up to order $\mathrm{O}(n^{-1})$ is

$$\begin{aligned} &E\big[n(\hat{\beta}_{\mathrm{FG}\iota\mathrm{S}} - \beta)(\hat{\beta}_{\mathrm{FG}\iota\mathrm{S}} - \beta)'\big] \\ &\quad = E(\xi_0\xi_0') + E(\xi_0\xi_{-1/2}' + \xi_{-1/2}\xi_0') + E(\xi_{-1/2}\xi_{-1/2}') \\ &\qquad + E(\xi_0\xi_{-1}' + \xi_{-1}\xi_0'), \end{aligned} \tag{3.37}$$

where from (3.32) we have

$$\begin{aligned} &E(\xi_0\xi_0') = n\sigma_u^2(X'\varOmega X)^{-1}, \\ &E(\xi_0\xi_{-1/2}') = \lambda A^{-1}X'\varOmega E(ww'f)P_1'A^{-1}/n, \\ &E(\xi_{-1/2}\xi_{-1/2}') = \lambda^2 A^{-1}P_1 E\big(ww'f^2\big)P_1'A^{-1}/n, \\ &E(\xi_0\xi_{-1}') = \lambda^2 A^{-1}\big(X'\varOmega/\sqrt{n}\big)E\big(ww'f^2\big)P_2'A^{-1}/n \\ &\qquad\qquad + \lambda A^{-1}\big(X'\varOmega/\sqrt{n}\big)E\big[\big(f^* - fv_v + f^2\big)ww'\big]P_1'A^{-1}/n. \end{aligned}$$

Consider the expectation

$$E(ww'f) = DE(f\alpha\alpha')D' + DE(f\alpha u') + E(fu\alpha')D' + E(fuu'), \tag{3.38}$$

where

$$\begin{aligned} E(f\alpha\alpha') &= E\big(u'\big(Q/(T-1) - \lambda\overline{Q}\big)u\sigma_u^{-2}/\sqrt{n}\big)E(\alpha\alpha') \\ &\qquad - (1-\lambda)\sigma_\alpha^{-2}E(\alpha'\alpha \cdot \alpha\alpha')/\sqrt{n} \\ &= -(1-\lambda)(2+\gamma_{2\alpha})\sigma_\alpha^2 I_n/\sqrt{n}, \end{aligned}$$

$$\begin{aligned}
E(f\alpha u') &= -2\sigma_\eta^{-2}E(\alpha\alpha')D'E(uu')/\sqrt{n}\\
&= -2\lambda\sigma_\alpha^2\sigma_\eta^{-2}D'/\sqrt{n},\\
E(fu\alpha') &= -2\lambda\sigma_\alpha^2\sigma_\eta^{-2}D/\sqrt{n},\\
E(fuu') &= \sigma_u^{-2}E\big[u'\big(Q/(T-1)-\lambda\overline{Q}\big)u\cdot uu'\big]/\sqrt{n}\\
&\quad -(1-\lambda)\sigma_\alpha^{-2}E(\alpha\alpha')E(uu')/\sqrt{n}\\
&= \sigma_u^2\big[(1-\lambda)\gamma_{2u}I_{nT}/T+2(T-1)Q-2\lambda\overline{Q}\big]/\sqrt{n}.
\end{aligned}$$

Now substitute these four terms into $E(ww'f)$, and we get

$$\begin{aligned}
E(ww'f) = \sigma_u^2\big[(1-\lambda)\big(\gamma_{2u}I_{nT}/T-(1-\lambda)\gamma_{2\alpha}\bar{Q}/\lambda\big)+2Q/(T-1)\\
-2\overline{Q}/\lambda\big]/\sqrt{n}.
\end{aligned} \tag{3.39}$$

Next let us define $Z_u = u/\sigma_u$, $Z_\alpha = \alpha/\sigma_\alpha$, and the first four moments of the elements of Z_u and Z_α are given in Appendix A3. Then

$$\begin{aligned}
E\big(ww'f^2\big) = DE\big(f^2\alpha\alpha'\big)D' + DE\big(f^2\alpha u'\big)\\
+E\big(f^2u\alpha'\big)D' + E\big(f^2uu'\big).
\end{aligned} \tag{3.40}$$

Consider the first term on the right-hand side of (3.40) we note that

$$\begin{aligned}
&E\big(\alpha\alpha' f^2\big)\\
&\quad= E\big[\alpha\alpha'\big(v_u^2+(1-\lambda)^2v_\alpha^2+\lambda^2\varepsilon_u^2+4v_{\alpha u}^2-2(1-\lambda)v_uv_\alpha\big)\big]\\
&\qquad+E\big[\alpha\alpha'\big(-2\lambda v_u\varepsilon_u-4v_uv_{\alpha u}+2\lambda(1-\lambda)v_\alpha\varepsilon_u\big)\big]\\
&\qquad+E\big[\alpha\alpha'\big(4(1-\lambda)v_\alpha v_{\alpha u}+4\lambda\varepsilon_uv_{\alpha u}\big)\big]\\
&\quad= \sigma_u^2I_nE\big(v_u^2+\lambda^2\varepsilon_u^2-2\lambda v_u\varepsilon_u\big)\\
&\qquad+E\big[\alpha\alpha'(-4v_uv_{\alpha u}+4\lambda\varepsilon_uv_{\alpha u})\big]\\
&\qquad+E\big[\alpha\alpha'\big(4v_{\alpha u}^2-2(1-\lambda)v_uv_\alpha+2\lambda(1-\lambda)v_\alpha\varepsilon_u\big)\big]\\
&\qquad+E\big[\alpha\alpha'(1-\lambda)^2v_\alpha^2\big]+4(1-\lambda)E[\alpha\alpha'v_\alpha v_{\alpha u}]\\
&\quad= I+II+III+IV+V,
\end{aligned} \tag{3.41}$$

where

$$\begin{aligned}
I &= \sigma_\alpha^2I_nE(v_u-\lambda\varepsilon_u)^2\\
&= n\sigma_\alpha^2I_nE\big[Z_u'\big(Q/(T-1)-\lambda\overline{Q}\big)Z_u/n-(1-\lambda)\big]^2\\
&= \sigma_\alpha^2I_n\big[\gamma_{2u}(1-\lambda)^2/T+2/(T-1)+2\lambda^2\big],\\
II &= -4E\big[\alpha\alpha'v_{\alpha u}(v_u-\lambda\varepsilon_u)\big]\\
&= -4\frac{\sigma_u}{\sigma_\eta^2}E_\alpha\big[\alpha\alpha'\cdot\alpha'D'E_{Z_u}\big(Z_u\cdot Z_u'\big(Q/(T-1)-\lambda\overline{Q}\big)Z_u/n\big)\big]
\end{aligned}$$

$$\begin{aligned}
&= -4(1-\lambda)\gamma_{1u}\gamma_{1\alpha}\sigma_u\sigma_\alpha^3\sigma_\eta^{-2}I_n/n = \mathrm{O}\bigl(n^{-1}\bigr),\\
III &= E\bigl[\alpha\alpha'\bigl(4v_{\alpha u}^2 - 2(1-\lambda)v_\alpha(v_u - \lambda\varepsilon_u)\bigr)\bigr]\\
&= E\bigl[\alpha\alpha'\bigl(4\bigl(u'D\alpha\sigma_\eta^{-2}/\sqrt{n}\bigr)^2 - 2(1-\lambda)n\bigl(\alpha\alpha'\sigma_\alpha^{-2}/n - 1\bigr)\\
&\qquad \times \bigl(u'\bigl(Q/T - 1(T-1) - \lambda\overline{Q}\bigr)u\sigma_u^{-2}/n - (1-\lambda)\bigr)\bigr)\bigr]\\
&= 4\lambda(1-\lambda)\sigma_\alpha^2 I_n + \mathrm{O}\bigl(n^{-1}\bigr),\\
IV &= E\bigl[\alpha\alpha'(1-\lambda)^2 n\bigl(\alpha'\alpha\sigma_\alpha^2/n - 1\bigr)^2\bigr]\\
&= (1-\lambda)^2\sigma_\alpha^2 E\bigl[Z_\alpha Z_\alpha'\bigl((Z_\alpha' Z_\alpha)^2/n - 2Z_\alpha Z_\alpha' + n\bigr)\bigr]\\
&= (1-\lambda)^2\sigma_\alpha^2(\gamma_{2\alpha} + 2)I_n,\\
V &= 0.
\end{aligned}$$

Substitute the above five results in (3.41), we have

$$DE\bigl(\alpha\alpha' f^2\bigr)D' = T\sigma_\alpha^2\bigl[(1-\lambda)^2(\gamma_{2u}/T + \gamma_{2\alpha}) + 2T/(T-1)\bigr]\overline{Q}. \tag{3.42}$$

In the second term on the right-hand side of (3.40)

$$\begin{aligned}
E\bigl(\alpha u' f^2\bigr) &= E\bigl[\alpha u'(v_u - \lambda\varepsilon_u)^2\bigr] - 4E\bigl[\alpha u' v_{\alpha u}(v_u - \lambda\varepsilon_u)\bigr]\\
&\quad + E\bigl[\alpha u'\bigl(4v_{\alpha u}^2 - 2(1-\lambda)v_\alpha(v_u - \lambda\varepsilon_u)\bigr)\bigr]\\
&\quad + E\bigl[\alpha u'(1-\lambda)^2 v_\alpha^2\bigr] + 4(1-\lambda)E(\alpha u' v_\alpha v_{\alpha u}),
\end{aligned} \tag{3.43}$$

where

$$\begin{aligned}
I &= 0,\\
II &= -4E\bigl[\alpha u' u' D\alpha\bigl(u'\bigl(Q/(T-1) - \lambda\overline{Q}\bigr)u\sigma_u^{-2}/n - (1-\lambda)\bigr)/\sigma_\eta^2\bigr]\\
&= -4\sigma_u^2\sigma_\alpha^2 D' E\bigl[Z_u Z_u'\bigl(Z_u'\bigl(Q/(T-1) - \lambda\overline{Q}\bigr)\\
&\qquad \times Z_u/n - (1-\lambda)\bigr)\bigr]/\sigma_\eta^2\\
&= -4\sigma_u^2\sigma_\alpha^2 D'\bigl[\gamma_{2u}(1-\lambda)I_{nT}n^{-1}T^{-1}\\
&\qquad + 2\bigl(Q/(T-1) - \lambda\overline{Q}\bigr)/n\bigr]/\sigma_\eta^2\\
&= \mathrm{O}\bigl(n^{-1}\bigr),\\
III &= E\bigl[\alpha u'\bigl(4u'D\alpha\alpha' D' u\sigma_\eta^2/n - 2(1-\lambda)n\bigl(\alpha\alpha'\sigma_\alpha^{-2}/n - 1\bigr)\\
&\qquad \times \bigl(u'\bigl(Q/(T-1) - \lambda\overline{Q}\bigr)u\sigma_u^{-2}/n - (1-\lambda)\bigr)\bigr)\bigr]\\
&= \mathrm{O}\bigl(n^{-1}\bigr),\\
IV &= 0,\\
V &= 4(1-\lambda)E\bigl[\alpha u'\bigl(\alpha\alpha'\sigma_\alpha^{-2}/n - 1\bigr)u' D\alpha\sigma_\eta^{-2}\bigr]\\
&= 4(1-\lambda)\sigma_u^2\sigma_\eta^{-2}E\bigl[\alpha\alpha'\bigl(\alpha\alpha'\sigma_\alpha^{-2}/n - 1\bigr)\bigr]D'
\end{aligned}$$

$$
\begin{aligned}
&= 4(1-\lambda)\sigma_u^2\sigma_\alpha^2\sigma_\eta^{-2}\big(\gamma_{2\alpha}n^{-1}T^{-1}+1-2/n-1\big)I_nD' \\
&= \mathrm{O}\big(n^{-1}\big).
\end{aligned}
$$

Substitute above five results into (3.43) and we get

$$
E\big(\alpha u' f^2\big) = \mathrm{O}\big(n^{-1}\big). \tag{3.44}
$$

The fourth term on the right-hand side of (3.40) is

$$
\begin{aligned}
E\big(f^2uu'\big) = {}& E\big[uu'(v_u-\lambda\varepsilon_u)^2\big] - 4E\big[uu'\big(v_{\alpha u}(v_u-\lambda\varepsilon_u)\big)\big] \\
&+ E\big[uu'\big(4v_{\alpha u}^2 - 2(1-\lambda)v_\alpha(v_u-\lambda\varepsilon_u)\big)\big] \\
&+ E\big[uu'\big((1-\lambda)^2v_\alpha^2\big)\big] + 4E\big[uu'\big((1-\lambda)v_\alpha v_{\alpha u}\big)\big] \\
= {}& I + II + III + IV + V,
\end{aligned} \tag{3.45}
$$

where

$$
\begin{aligned}
I &= E\big[uu'n\big(u'\big(Q/(T-1)-\lambda\overline{Q}\big)u\sigma_u^{-2}/n-(1-\lambda)\big)^2\big] \\
&= \sigma_u^2E\big[n^{-1}Z_uZ_u'\cdot\big(\big(Z_u'\big(Q/(T-1)-\lambda\overline{Q}\big)Z_u\big)^2 \\
&\qquad - 2n(1-\lambda)Z_u'\big(Q/(T-1)-\lambda\overline{Q}\big)Z_u + n^2(1-\lambda)^2\big)\big] \\
&= \sigma_u^2\big[(1-\lambda)^2\gamma_{2u}/T + 2\big(1/(T-1)+\lambda^2\big)\big]I_{nT}, \\
II &= 0, \\
III &= 4\sigma_\alpha^2\sigma_\eta^{-4}E(uu'\cdot u'DD'u)/n \\
&= 4\sigma_u^4\sigma_\alpha^2\sigma_\eta^{-4}TI_{nT} \\
&= 4(1-\lambda)\lambda\sigma_u^2I_{nT} + \mathrm{O}\big(n^{-1}\big), \\
IV &= (1-\lambda)^2E\big[uu'n\big(\alpha\alpha'\sigma_\alpha^{-2}/n-1\big)^2\big] \\
&= \sigma_u^2(1-\lambda)^2nE\big(\alpha\alpha'\sigma_\alpha^{-2}/n-1\big)^2I_{nT} \\
&= \sigma_u^2(1-\lambda)^2(\gamma_{2\alpha}+2)I_{nT}, \\
V &= 4(1-\lambda)E\big[uu'u'D\alpha\sigma_\eta^{-2}\big(\alpha\alpha'\sigma_\alpha^{-2}/n-1\big)\big] \\
&= 4(1-\lambda)\sigma_\alpha^{-2}\sigma_\eta^{-2}E(uu'\cdot u'D\alpha\cdot\alpha'\alpha)/n = \mathrm{O}\big(n^{-1}\big).
\end{aligned}
$$

Hence

$$
E\big(uu'f^2\big) = \sigma_u^2\big[(1-\lambda)^2(\gamma_{2u}/T+\gamma_{2\alpha}) + 2T/(T-1)\big]I_{nT}, \tag{3.46}
$$

and (3.40) can be written as

$$
\begin{aligned}
E\big(ww'f^2\big) = \sigma_u^2\big[&(1-\lambda)^2(\gamma_{2u}/T+\gamma_{2\alpha}) \\
&+ 2T/(T-1)\big]\left(\frac{1-\lambda}{\lambda}\overline{Q} + I_{nT}\right).
\end{aligned} \tag{3.47}
$$

Next, let us consider $E(f^*ww')$ in (3.37)

$$\begin{aligned} E(f^*ww') &= E(f^*uu') + DE(f^*\alpha u') + E(f^*u\alpha')D' \\ &\qquad + DE(f^*\alpha\alpha')D' = I + II + III + IV, \end{aligned} \tag{3.48}$$

where

$$\begin{aligned} I &= E\big[\big(v_\alpha^* + \varepsilon_u^* + 2v_{\alpha u}^* - v_u^*/(T-1) - k(T-2)/(T-1)\big)uu'\big] \\ &= \sigma_u^2\big[\sigma_\alpha^2\sigma_\eta^{-2}kT - k(T-2)/(T-1)\big] \\ &\quad + E\big[\big(\varepsilon_u^* - v_u^*/(T-1)\big)uu'\big] \\ &= \sigma_u^2\gamma_{2u}\big[\big(\lambda\big(I_{nT} \odot \bar{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q}\big) \\ &\quad - \big(I_{nT} \odot QX(X'QX)^{-1}X'Q\big)/(T-1)\big)\big] \\ &\quad + \sigma_u^2\big[2\big(\lambda\overline{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q} - QX(X'QX)^{-1}X'Q/(T-1)\big)\big], \\ II &= DE[2v_{\alpha u}^* \cdot \alpha u'] \\ &= 2DE\big[\alpha\alpha'D'X(X'\overline{Q}X)^{-1}X'\overline{Q}uu'\sigma_\eta^{-2}\big] \\ &= 2(1-\lambda)\sigma_u^2\overline{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q} = III, \\ IV &= DE\big[\big(v_\alpha^* + \varepsilon_u^* + 2v_{\alpha u}^* - v_u^*(T-1) - k(T-2)/(T-1)\big)\alpha\alpha'\big]D' \\ &= DE[v_\alpha^*\alpha\alpha']D' + \sigma_\alpha^2DD'\big(k\lambda - k/(T-1) - k(T-2)/(T-1)\big) \\ &= (1-\lambda)^2\lambda^{-1}\gamma_{2\alpha}D\big(I_n \odot D'X(X'\overline{Q}X)^{-1}X'D\big)D'/T^2 \\ &\quad + 2(1-\lambda)^2\lambda^{-1}\sigma_u^2\overline{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q}. \end{aligned}$$

Thus

$$\begin{aligned} E(f^*ww') = \sigma_u^2\big[&\gamma_{2u}\big(\lambda\big(I_{nT} \odot \overline{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q}\big) \\ &- \big(I_{nT} \odot QX(X'QX)^{-1}X'Q\big)/(T-1)\big)\big] \\ &+ \sigma_u^2\big[(1-\lambda)^2\lambda^{-1}T^{-2}\gamma_{2\alpha} \\ &\times D\big(I_n \odot D'X(X'\bar{Q}X)^{-1}X'D\big)D' \\ &- 2QX(X'QX)^{-1}X'Q/(T-1)\big] \\ &+ \sigma_u^2\big[2\lambda^{-1}\overline{Q}X(X'\overline{Q}X)^{-1}X'\overline{Q}\big]. \end{aligned} \tag{3.49}$$

Consider $E(fv_uww')$ in (3.37)

$$\begin{aligned} E(fv_uww') &= E(fv_uuu') + E(fv_uu\alpha')D' + DE(fv_u\alpha u') \\ &\qquad + DE(fv_u\alpha\alpha')D' = I + II + III + IV, \end{aligned} \tag{3.50}$$

where

$$\begin{aligned} I &= E\big[(v_u - \lambda\varepsilon_u)v_u \cdot uu'\big] \\ &= n\sigma_u^2E\big[\big(Z_u'\big(Q/\big(n(T-1)\big) - \lambda\overline{Q}/n\big)Z_u - (1-\lambda)\big) \end{aligned}$$

$$\begin{aligned}
&\quad\times \big(Z_u' Q Z_u n^{-1}/(T-1) - 1\big) Z_u' Z_u\big] \\
&= \sigma_u^2\big[2/(T-1) + \gamma_{2u}(1-\lambda)/T\big] I_{nT}, \\
II &= E\big[\big(v_u - (1-\lambda)v_\alpha - \lambda\varepsilon_u - 2v_{\alpha u}\big)v_u \cdot u\alpha'\big]D' \\
&= E\big[-(1-\lambda)v_\alpha v_u \cdot u\alpha'\big]D' + E(-2v_{\alpha u}v_u \cdot u\alpha')D' \\
&= -(1-\lambda)E(v_u u)E(v_\alpha \alpha')D' - 2\sigma_\alpha^2\sigma_\eta^{-2}E(v_u u u')DD'/\sqrt{n} \\
&= \mathrm{O}\big(n^{-1}\big) = III, \\
IV &= DE\big[\big(v_u - (1-\lambda)v_\alpha - \lambda\varepsilon_u - 2v_{\alpha u}\big)v_u \alpha\alpha'\big]D' \\
&= \sigma_\alpha^2 DE\big[(v_u - \lambda\varepsilon_u)v_u\big]D' - 2DE(v_{\alpha u}v_u\alpha\alpha')D' \\
&= \sigma_u^2(1-\lambda)\lambda^{-1}\big[\gamma_{2u}(1-\lambda)T^{-1} + 2/(T-1)\big]\overline{Q} + \mathrm{O}\big(n^{-1}\big).
\end{aligned}$$

Therefore

$$E(f v_u w w') = \sigma_u^2\lambda^{-1}\big[\gamma_{2u}(1-\lambda)/T + 2/(T-1)\big](\overline{Q} + \lambda Q). \quad (3.51)$$

Plugging (3.39), (3.47), (3.49), and (3.51) into (3.37) we have

$$\begin{aligned}
E(\xi_0\xi_{-1/2}') &= \frac{\lambda}{n}A^{-1}X'\Omega E(ww'f)P_1'A^{-1} \\
&= -\lambda\sigma_u^2\big[\gamma_{2u}(1-\lambda)^2T^{-1} + \gamma_{2u}(1-\lambda)^2T^{-1} \\
&\quad + 2T/(T-1)\big]\Delta/n = E(\xi_{-1/2}\xi_0'), \\
E(\xi_{-1/2}\xi_{-1/2}') &= \lambda^2A^{-1}P_1E\big(ww'f^2\big)P_1'A^{-1}/n \\
&= \lambda^2\sigma_u^2\big[(1-\lambda)^2\big(\gamma_{2u}T^{-1} + \gamma_{2\alpha}\big) + 2T/(T-1)\big]\Delta/n, \\
E(\xi_0\xi_{-1}') &= C + \frac{2\lambda\sigma_u^2T}{n(T-1)}\Delta, \\
E(\xi_{-1}\xi_0') &= C' + \frac{2\lambda\sigma_u^2T}{n(T-1)}\Delta.
\end{aligned}$$

Using these in (3.37) the MSE result in Theorem 3.1 follows. □

3.4 Numerical results

In this section we provide a numerical study of the behavior of analytical Bias and MSE under non-normality. The data generating process is specified as follows

$$y_{it} = x_{it}\beta + \alpha_i + u_{it}.$$

x_{it} are generated via the method of Nerlove (1971)

$$x_{it} = 0.1t + 0.5x_{it-1} + w_{it},$$

Table 3.3. ***$n = 10$, $T = 50$, α is non-normal and u is normal, $\sigma_\alpha = 1$, $\sigma_u = 0.6$***

θ_2	$\gamma_{1\alpha}$	$\gamma_{2\alpha}$	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	0.001297	0	0.000098	0.000100
2.0	−1.70	5.64	0.000833	0	0.000099	0.000100
2.5	−1.23	2.93	0.000603	0	0.000100	0.000100
3.0	−0.94	1.74	0.000461	0	0.000100	0.000100

$$x_{i0} = 10 + 5w_{i0},$$
$$w_{it} \sim U\left[-\frac{1}{2}, \frac{1}{2}\right].$$

We omit the constant term and consider the data generating process described in Corollary 3.2 and Corollary 3.3. For Corollary 3.2, we let $\beta = 0.5$. $u_{it} \sim IIN(0, 0.36)$, which implies $\gamma_{1u} = \gamma_{2u} = 0$. α_i are generated by Johnson's (1949) S_u system, introducing non-normality to our data generating process. The non-normal α_i is generated by transforming a standard normal random variable ε_i

$$\alpha_i^* = \sinh\left(\frac{\varepsilon_i - \theta_1}{\theta_2}\right),$$

and letting α_i be the standardized version of α_i^* with zero mean and variance is one.

Different values of (θ_1, θ_2) givens different values of the skewness and kurtosis of the random variable α_i^*. Define $\omega = \exp(\theta_2^{-2})$ and $\psi = \theta_1/\theta_2$ and the four moments of α_i are given by

$$\begin{aligned}
E(\alpha_i^*) &= \mu_\alpha = -\omega^{1/2}\sinh(\psi),\\
E(\alpha_i^* - \mu_\alpha)^2 &= \frac{1}{2}(\omega - 1)\big[\omega\cosh(2\psi) + 1\big],\\
E(\alpha_i^* - \mu_\alpha)^3 &= -\frac{1}{4}\omega^{1/2}(\omega - 1)^2\big[\omega(\omega + 2)\sinh(3\psi) + 3\sinh(\psi)\big],\\
E(\alpha_i^* - \mu_\alpha)^4 &= \frac{1}{8}(\omega - 1)^2\big[\omega^2\big(\omega^4 + 2\omega^3 + 3\omega^2 - 3\big)\cosh(4\psi)\\
&\quad + 4\omega^2(\omega + 2)\cosh(2\psi) + 3(2\omega + 1)\big].
\end{aligned}$$

From this we get skewness $\gamma_{1\alpha} = E(\alpha_i^* - \mu_\alpha)^3/(E(\alpha_i^* - \mu_\alpha)^2)^{3/2}$ and excess kurtosis $\gamma_{2\alpha} = E(\alpha_i^* - \mu_\alpha)^4/(E(\alpha_i^* - \mu_\alpha)^2)^2 - 3$. In Tables 3.1 to 3.3, θ_1 is set to be 4 and $\theta_2 \in [1.5, 3]$. This combination of θ_1 and θ_2 gives a moderate interval for the variance of α_i^*, from 0.5 to 45. For Corollary 3.3, we apply the same method to the generation of non-normal u_{it},

Table 3.4. $\boldsymbol{n = 10, T = 5}$***, a is normal and u is non-normal,*** $\boldsymbol{\sigma_\alpha = 2, \sigma_u = 1}$

θ_2	γ_{1u}	γ_{2u}	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	−0.001906	0	0.007143	0.007135
2.0	−1.70	5.64	−0.001224	0	0.007243	0.007135
2.5	−1.23	2.93	−0.000886	0	0.007273	0.007135
3.0	−0.94	1.74	−0.000677	0	0.007287	0.007135

and let $\alpha_i \backsim IIN(0, 4)$. In order to investigate the finite sample behavior of Bias and MSE, we let $n = 10$ and $T = 5$. We replicate the experiment 1000 times for each pair of (θ_1, θ_2).

When only α is non-normal, we note that from Table 3.1 that the MSE changes with $\gamma_{2\alpha}$. Generally, for some large $\gamma_{2\alpha}$, approximate MSE is less than asymptotic MSE while for some small $\gamma_{2\alpha}$, approximate MSE is greater than asymptotic MSE. Thus the use of the asymptotic MSE, when the sample is small or moderately large, will provide an under estimation or over estimation depending on the magnitude of $\gamma_{2\alpha}$. Further the t-ratios for hypothesis testing, based on asymptotic MSE, may provide under or over rejection of the null hypothesis. When the sample is moderately large (Table 3.2) we get similar results, but the asymptotic MSE is the same as the approximate MSE up to 4 digits. However, for the cases when only u_{it} is non-normal we see from Table 3.4 that the approximate MSE is greater than the asymptotic MSE for all values of γ_{2u}. Thus, in this case, the use of asymptotic MSE in practice, will generally provide underestimation of MSE and t-ratios may falsely reject the null hypothesis. For moderately large samples in Table 3.5, the approximate MSE is still greater than the asymptotic MSE, but they are the same up to 4 digits. Thus, when either alpha or u is non-normally distributed, we observe that while the use of the asymptotic MSE may provide under or over estimation of the MSE, the asymptotic MSE estimates the approximate MSE accurately since they are the same up to three or four digits, especially for moderately large samples.

In Remark 3.1, Bias is found to be a decreasing function of $\gamma_{1\alpha}$ and an increasing function of γ_{1u}, which is consistent with the results seen in Tables 3.1–3.6. The monotonic relations between Bias and variances of the error components in Remark 3.2 are shown numerically in Tables 3.8–3.11, where in Tables 3.8–3.9 we fix σ_u and increase σ_α and in Tables 3.10–3.11 we fix σ_α and increase σ_u.

We also simulate the different ns for the same T and vice versa. The results presented here are for $T = 5$ with $n = 10{,}50$ and for $n = 10$

Table 3.5. ***$n = 50$, $T = 5$, α is normal and u is non-normal, $\sigma_\alpha = 2$, $\sigma_u = 1$***

θ_2	γ_{1u}	γ_{2u}	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	−0.000397	0	0.001553	0.001553
2.0	−1.70	5.64	−0.000255	0	0.001558	0.001553
2.5	−1.23	2.93	−0.000185	0	0.001559	0.001553
3.0	−0.94	1.74	−0.000141	0	0.001560	0.001553

Table 3.6. ***$n = 10$, $T = 50$, α is normal and u is non-normal, $\sigma_\alpha = 2$, $\sigma_u = 1$***

θ_2	γ_{1u}	γ_{1u}	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	−0.000018	0	0.000281	0.000281
2.0	−1.70	5.64	−0.000012	0	0.000282	0.000281
2.5	−1.23	2.93	−0.000009	0	0.000282	0.000281
3.0	−0.94	1.74	−0.000007	0	0.000282	0.000281

Table 3.7. ***$n = 10$, $T = 5$. Both α and u are normal, $\sigma_u = 0.6$***

σ_α	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1	0	0	0.002612	0.002535
5	0	0	0.002936	0.002931
10	0	0	0.002948	0.002947
15	0	0	0.002950	0.002950

Table 3.8. ***$n = 10$, $T = 5$, α is normal and u is non-normal, $\sigma_\alpha = 0.5$, $\sigma_u = 2$, $\lambda = 0.76$***

θ_2	$\gamma_{1\alpha}$	$\gamma_{2\alpha}$	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	0.002723	0	0.012439	0.011967
2.0	−1.70	5.64	0.001748	0	0.012542	0.011967
2.5	−1.23	2.93	0.001266	0	0.012573	0.011967
3.0	−0.94	1.74	0.000968	0	0.012587	0.011967

with $T = 5, 50$. The results for other values of n and T are available from the authors upon request, and they give the similar conclusions. When α is non-normal, the maximum relative bias, $E(\hat{\beta} - \beta)/\beta$, decreases from 2.7% to 0.5% when n changes from 10 to 50 with $T = 5$; and it decreases

Table 3.9. **$n = 10, T = 5$, *α is non-normal and u is normal,* $\sigma_\alpha = 1.5, \sigma_u = 2, \lambda = 0.26$**

θ_2	$\gamma_{1\alpha}$	$\gamma_{2\alpha}$	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	0.029105	0	0.019418	0.021358
2.0	−1.70	5.64	0.018689	0	0.021402	0.021358
2.5	−1.23	2.93	0.013533	0	0.021998	0.021358
3.0	−0.94	1.74	0.010343	0	0.022260	0.021358

Table 3.10. **$n = 10, T = 5$, *α is normal and u is non-normal,* $\sigma_\alpha = 2, \sigma_u = 10, \lambda = 0.83$**

θ_2	γ_{1u}	γ_{2u}	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	−0.007749	0	0.298761	0.282281
2.0	−1.70	5.64	−0.004976	0	0.297935	0.282281
2.5	−1.23	2.93	−0.003603	0	0.297687	0.282281
3.0	−0.94	1.74	−0.002754	0	0.297578	0.282281

Table 3.11. **$n = 10, T = 5$, *α is normal and u is non-normal,* $\sigma_\alpha = 2, \sigma_u = 20, \lambda = 0.95$**

θ_2	γ_{1u}	γ_{2u}	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	−0.004111	0	0.994565	0.939214
2.0	−1.70	5.64	−0.002640	0	0.990324	0.939214
2.5	−1.23	2.93	−0.001912	0	0.989049	0.939214
3.0	−0.94	1.74	−0.001461	0	0.988490	0.939214

Table 3.12. **$n = 10, T = 5$. *Both α and u are non-normal,* $\sigma_\alpha = 5, \sigma_u = 10$**

θ_2	γ_{1u}	γ_{2u}	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	0.043545	0	0.349311	0.365864
2.0	−1.70	5.64	0.027961	0	0.373554	0.365864
2.5	−1.23	2.93	0.020247	0	0.380839	0.365864
3.0	−0.94	1.74	0.015474	0	0.384039	0.365864

from 2.7% to 0.3% when $n = 10$ and T changes from 5 to 50. When u is non-normal, the maximum relative bias changes from 0.4% to 0.08% for the change of n from 10 to 50 with $T = 5$; and from 0.4% to 0.004%

Table 3.13. ***$n = 10$, $T = 5$. Both α and u are non-normal, $\sigma_\alpha = 10$, $\sigma_u = 5$***

θ_2	γ_{1u}	γ_{2u}	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	0.081264	0	0.152518	0.168328
2.0	−1.70	5.64	0.052182	0	0.164591	0.168328
2.5	−1.23	2.93	0.037786	0	0.168220	0.168328
3.0	−0.94	1.74	0.028878	0	0.169814	0.168328

Table 3.14. ***$n = 10$, $T = 5$. Both α and u are non-normal, $\sigma_\alpha = 10$, $\sigma_u = 10$***

θ_2	γ_{1u}	γ_{2u}	Approx Bias	Asym Bias	Approx MSE	Asym MSE
1.5	−2.65	14.65	0.141105	0	0.457876	0.537747
2.0	−1.70	5.64	0.090608	0	0.523752	0.537747
2.5	−1.23	2.93	0.065611	0	0.543550	0.537747
3.0	−0.94	1.74	0.050143	0	0.552245	0.537747

for $n = 10$ when T changing from 5 to 50. Thus the order of bias is not very significant, further, it is found that for a fixed T, e.g., $T = 5$, when n is large enough, for example, 50, the approximate bias is practically zero. These results are consistent with the results in Maddala and Mount (1973). For the MSE, when α is non-normal and T is fixed at 5, the approximate MSE is equal to asymptotic MSE up to the third digit when $n = 10$, but up to the fourth digit when $n = 50$. For the case when n is fixed at 10 and T changes from 5 to 50, the two MSEs are the same up to three digits. Similar results hold for the case when u is non-normally distributed.

Next we consider the DGPs with both error components are non-normally distributed and have large variances in small sample. It is found in such cases the relative bias can be large and asymptotic MSE may not be very accurate. Tables 3.12–3.14 give some examples. Most tables show that the approximate bias is not negligible. The range of relative bias in Table 3.12 is [3%, 8.7%] and it increases to [10%, 28%] in Table 3.14. The approximate and asymptotic MSEs can be different even at the first digit, as shown in the first row of Table 3.14.

In Table 3.7, both α and u are normal, where $u_{it} \sim IIN(0, 0.36)$ and α_i has zero mean and changing variance. $\gamma_{1\alpha} = \gamma_{2\alpha} = \gamma_{1u} = \gamma_{2u} = 0$. In this case, the approximate MSE is always larger than asymptotic MSE, and this is consistent with the results in Corollary 3.1 and Remark 3.3. However, the difference in the approximate and asymptotic MSEs is the same up to 5 digits.

3.5 Conclusion

In this paper, we study the finite sample properties of the FGLS estimators in random-effects model with non-normal errors. We derive the asymptotic expansion of the Bias and MSE up to $O(n^{-1})$ and $O(n^{-2})$, respectively.

Firstly, the Bias depends only on skewness coefficient. Bias is zero for symmetric distributions or for distributions satisfying $\gamma_{1u}/\gamma_{1\alpha} = T\sigma_{\alpha}/\sigma_{u}$. We find Bias is a non-decreasing function of γ_{1u} and a non-increasing function of $\gamma_{1\alpha}$ provided $X'\iota_{nT} \geq 0$. Under certain parameter restrictions, Bias is also found to be monotonic functions of variances of the error components.

Secondly, the MSE depends only on the kurtosis coefficient. The approximate MSE can be greater or smaller than asymptotic MSE. The statistical inference based on using the asymptotic MSE can be quite accurate when variances of the error components are small since it is the same as the approximate MSE, under the normality as well as a non-normal distribution considered, up to three or four digits, especially for moderately large samples. However, when those variances are large, asymptotic results can give inaccurate results.

Acknowledgements

The authors dedicate this work in memory of the late Viren Srivastava with whom the initial work on this paper began. We are thankful to Badi Baltagi and a referee for their useful comments on this paper. We gratefully acknowledge helpful suggestions from A. Chaturvedi, S. Kumar and K. Mundra. We also thank seminar participants at the 11th world conference on panel data for helpful comments. The first author also acknowledges the financial support from the Academic Senate, U.C.R. and the second author thanks the Department of Economics, U.C.R. for financial support. All remaining errors are our own.

Appendix A3

The following results have been repeatedly used in the derivation in Section 3.3:

Let G_1 and G_2 be two $nT \times nT$ idempotent matrices with non-stochastic elements such that

$$\mathrm{tr}(G_1) = ng_1,$$

$$\mathrm{tr}(G_2) = ng_2,$$
$$\mathrm{tr}(G_{12}) = ng_{12}.$$

Assuming G_1 and G_2 to be symmetric matrices and Z to be an $nT \times 1$ random vector whose elements are i.i.d. with the first four moments given as[1]

$$Ez_j = 0, \qquad Ez_j^2 = 1, \qquad Ez_j^3 = \gamma_{1z},$$
$$Ez_j^4 = \gamma_{2z} + 3, \quad j = 1, \dots, nT.$$

Then we have

$$E(Z'G_1Z \cdot Z) = \gamma_{1z}(I_{nT} \odot G_1)\iota_{nT}, \tag{A3.1}$$
$$E(Z'G_1Z \cdot ZZ') = \gamma_{2z}(I_{nT} \odot G_1) + \mathrm{tr}(G_1) + 2G_1. \tag{A3.2}$$

Further, if the diagonal elements of G_1 are equal and those of G_2 are also equal, we have

$$E(Z'G_1Z \cdot Z) = \frac{\gamma_{1z}g_1}{T}\iota_{nT}, \tag{A3.3}$$
$$E(Z'G_1Z \cdot ZZ') = \frac{\gamma_{2z}g_1}{T}I_{nT} + ng_1I_{nT} + 2G_1, \tag{A3.4}$$

$$\begin{aligned}\frac{1}{n}&E(Z'G_1Z \cdot Z'G_2Z \cdot ZZ')\\ &= ng_1g_2I_{nT} + 2g_{12}I_{nT} + 2g_1G_2\\ &\quad + 2g_2G_1 + \frac{3g_1g_2\gamma_{2z}}{T}I_{nT} + \mathrm{O}\big(n^{-1}\big).\end{aligned} \tag{A3.5}$$

Notice that results (A3.1) to (A3.4) are exact while the result (A3.5) is given up to order $\mathrm{O}(n^{-1})$ only as it suffices for the present purpose.

References

Arellano, M., Bond, S. (1991), "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations", *Review of Economic Studies*, Vol. 58, pp. 277–297.

Baltagi, B.H. (1981), "Pooling: an experimental study of alternative testing and estimation procedures in a two-way error components model", *Journal of Econometrics*, Vol. 17, pp. 21–49.

Baltagi, B.H. (2001), *Econometric Analysis of Panel Data*, John Wiley & Sons, New York.

Fuller, W.A., Battese, G.E. (1974), "Estimation of linear models with cross-error structure", *Journal of Econometrics*, Vol. 2, pp. 67–78.

[1] If $Z = Z_\alpha$, the dimension changes to n, which implies $T = 1$ in the following results.

Johnson, N.L. (1949), “System of frequency curves generated by methods of transformation”, *Biometrika*, Vol. 36, pp. 149–176.

Kiviet, J.F. (1995), “On bias, inconsistency, and efficiency of various estimators in dynamic panel data models”, *Journal of Econometrics*, Vol. 68, pp. 53–78.

Maddala, G.S., Mount, T.D. (1973), “A comparative study of alternative estimators for variance components models used in econometric applications”, *Journal of the American Statistical Association*, Vol. 68, pp. 324–328.

Nerlove, M. (1967), “Experimental evidence on the estimation of dynamic economic relations from a time series of cross-sections”, *Economic Studies Quarterly*, Vol. 18, pp. 42–74.

Nerlove, M. (1971), “Further evidence on the estimation of dynamic economic relations from a time series of cross sections”, *Econometrica*, Vol. 39, pp. 358–382.

Taylor, W.E. (1980), “Small sample considerations in estimation from panel data”, *Journal of Econometrics*, Vol. 13, pp. 203–223.

This page intentionally left blank

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74004-4

CHAPTER 4

Modelling the Initial Conditions in Dynamic Regression Models of Panel Data with Random Effects

I. Kazemi[a] and R. Crouchley[b]

[a]Centre for Applied Statistics, Lancaster University, Lancaster, England
E-mail address: i.kazemi@lancaster.ac.uk
[b]Centre for e-Science, Lancaster University, Lancaster, England
E-mail address: r.crouchley@lancaster.ac.uk

Abstract

This paper compares dynamic panel regression with random effects under different assumptions about the nature of the initial conditions, and suggests that a pragmatic approach is to be preferred. The proposed approach has a flexible reduced form for the initial response which effectively imposes a random effect correlated with the subsequent model equation to deal with the initial conditions and to handle the problem of negative estimates of variance components. These concepts are illustrated by testing a variety of different hypothetical models in economic contexts. We use information criteria to select the best approximating model. We find that the full maximum likelihood improves the consistency results if the relationships between random effects, initial conditions and explanatory variables are correctly specified.

Keywords: dynamic panel data, random effects, initial conditions, non-negative estimates of variance components, MLEs

JEL classifications: C33, C59, O10

4.1. Introduction

Modelling dynamic regression for panel data has become increasingly popular in a wide variety of research areas over the past few decades. These models are specifically adapted for the statistical analysis of data that have a serial processes structure which allows for individual heterogeneity to control for time-invariant characteristics (Hausman and Taylor, 1981; Heckman, 1991) and dynamic feedback to make it easier for researchers to examine state dependence (Sevestre and Trognon, 1996; Nerlove, 1999). These two important ideas are often addressed by including individual-specific effects and a lagged endogenous variable in the regression model. A great deal of attention has been devoted to the problems created by these features with a particular focus on properties of different modelling strategies for the analysis of panel data.

The classical approach in the panel data literature is the use of fixed effects that simply ignores the component nature of residual heterogeneity. This will result in inconsistent estimates due to the problem of incidental parameters (Heckman, 1981; Lancaster, 2000) associated with the individual effects. The random effects model has been implemented to overcome the problem and consequently allows control of the unobserved effects by partitioning residual heterogeneity according to the within- and between-individual variations that exist in the data. A comprehensive review of the literature on the analysis of panel data with random effects is given by Nerlove (2002) and Hsiao (2002).

The main difficulty in the estimation of random effects is accounting for the initial conditions problem which arises if state dependence is suspected (Crouchley and Davies, 2001). This leads to a non-standard likelihood function which is not, in general, fully specified (Aitkin and Alfò, 2003) and to inconsistent ML estimates in the dynamic model (Anderson and Hsiao, 1982). For these reasons, a realistic solution is to treat the initial condition explicitly as endogenous and then use the likelihood of all observed outcomes including that in the initial time period. The likelihood approach then takes into account any information on the initial conditions in estimating the consistent regression parameters and likelihood equations for the components of covariance matrices. To do this, some special assumptions are required regarding the joint distribution of the first state on each individual and heterogeneity effects. A further factor in modelling initial conditions concerns the pre-sample history of the process generating the explanatory variables, which is usually unobservable and requires making additional assumptions about the data processes.

In the literature, the usual suggestion to overcome these problems is to suppose stationarity of the data processes (e.g., Bhargava and Sargan, 1983; Nerlove, 2000; Hsiao *et al.*, 2002) for both the response and explanatory variables. Although this hypothesis for the initial conditions is unlikely to lead always to the best models fitting the true processes, as we show in this paper, it has been used rather extensively by previous investigators without making any further effort to test the statistical hypothesis.

We specifically test the hypothetical models by making different alternative assumptions about the initial conditions: treating initial conditions as exogenous, in equilibrium, and adopting a set of flexible reduced forms.

Another important issue in the analysis of panel data using the classical ML approach for random effects is the occurrence of negative estimates of essentially non-negative parameters. Surprisingly, the variance estimate of random effects can take a zero or negative value. The usual suggestion for negative variance estimates (e.g., Breusch, 1987) is to replace these values by zero and refit the model. We show that the issue is completely different in the case of models which include initial conditions. Specifically, MLEs are obtained by finding the values of the parameters that maximise the likelihood function subject to the constraints that these parameters are within a known parameter space. We follow those algorithms with inequality constraints for solving the maximisation problem that involves finding estimates that satisfy the Kuhn–Tucker conditions (e.g., Driscoll and William, 1996).

Although there are many popular opinions about the initial conditions problem, little empirical analysis is available concerning its operation. There is a large empirical literature, for example, on classical dynamic growth studies (e.g., Knight *et al.*, 1993; Islam, 1995) that are typically carried out on fixed effects models. An exceptional study is Nerlove (2000) who estimates the model with random effects and addresses the problem by using conditional and unconditional likelihood. Following his approach, we further highlight the drawbacks of the ML approach using, as an illustration, economic growth models. Special attention is given to the properties of various models considered for initial conditions.

The organisation of this paper is as follows. Section 4.2 introduces the model of interest with random effects. Section 4.3 discusses the likelihood and the initial conditions problem. In Section 4.4 we briefly review the inconsistency of MLEs and the problem of estimating a negative value for the variance. Section 4.5 explains the full likelihood approach. Section 4.6 provides an overview of the historical development of knowledge about properties of modelling initial conditions and also introduces a pragmatic approach. Finally, the paper includes results from an empirical study, followed by model selection and recommendations.

4.2. The model with random effects

Suppose that Y_{it} is the response variable for individual i at time period t, while $\mathbf{X}_{it}$ is a $K \times 1$ vector of explanatory variables. Consider the following regression model for panel data:

$$\begin{aligned} &Y_{it} = \lambda + \gamma Y_{i,t-1} + \mathbf{X}'_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \\ &\quad i = 1, \ldots, N; t = 1, \ldots, T, \end{aligned} \tag{4.1}$$

where $Y_{i,t-1}$ is a lagged endogenous response variable and λ, γ, and $\boldsymbol{\beta}$ are regression coefficients.

Specifying the model with the lagged response has a significant advantage which derives from the fact that $Y_{i,t-1}$ summarises all the past effects of unmeasured variables on Y_{it}. This means, not only that the effects of measured explanatory variables $X_{it1}, \ldots, X_{itK}$ on Y_{it} can be estimated more accurately, but also that the coefficient on $Y_{i,t-1}$, γ, measures the effect of experience of the event one period ago on current values of Y_{it}. A positive value of γ indicates positive state dependence.

Adopting a conventional random effects approach, the usual assumptions are that the individual random effects $\alpha_i \sim \text{i.i.d.}(0, \sigma_\alpha^2)$; the unobserved time-varying errors $\varepsilon_{it} \sim \text{i.i.d.}(0, \sigma_\varepsilon^2)$; the α_i and the ε_{it} are independent for all i and t, and the stochastic variables $\mathbf{X}_{it}$ are strictly exogenous with respect to α_i and ε_{it}: $\text{cov}(\alpha_i, \mathbf{X}_{it}) = \mathbf{0}$, $\text{cov}(\varepsilon_{it}, \mathbf{X}_{jt}) = \mathbf{0}$ for all i, j and t.

4.3. The likelihood and initial conditions

The likelihood contribution of individual i is calculated by integrating over all possible values α_i given by

$$L_i(\boldsymbol{\phi}) = \int_{-\infty}^{\infty} f(Y_{i1}, \ldots, Y_{iT} | Y_{i0}, \alpha_i; \boldsymbol{\phi}) f(Y_{i0} | \alpha_i; \boldsymbol{\phi}) \, \mathrm{d}F(\alpha_i), \tag{4.2}$$

where $\boldsymbol{\phi}$ denotes a set of unknown model parameters, $F(\alpha_i)$ is the distribution function of α_i and $f(Y_{i0}|\alpha_i; \boldsymbol{\phi})$ refers to the marginal density of Y_{i0}, given α_i. The full likelihood function $L(\boldsymbol{\phi}) = \prod_i L_i(\boldsymbol{\phi})$ combines the conditional likelihood for each observation with the likelihood from the initial observations. We need only to integrate out the heterogeneity effects, α_i, by specifying the distribution of such effects and then maximising the likelihood function. The estimation of parameters $\boldsymbol{\phi}$ based on the full likelihood $L(\boldsymbol{\phi})$ introduces the question of the appropriate treatment of the initial conditions. More specifically, the difficulty is created by $f(Y_{i0}|\alpha_i; \boldsymbol{\phi})$ which cannot usually be determined without making additional assumptions. The naïve approach of treating the initial state Y_{i0} as exogenous and simply ignoring the initial conditions (Heckman, 1981) is refutable since the independence of Y_{i0} and unobserved heterogeneity effects, α_i, is highly implausible where modelling serial processes. Specifically, this restriction is appropriate only if the first period of observation of the serial process is the true initial period, or, if the true residuals that generate the stochastic process are serially independent. Treating the Y_{i0} as exogenous, the conditional probability of the initial sample state is $f(Y_{i0}|\alpha_i) = f(Y_{i0})$ and thus locates this term outside the integral (4.2). The likelihood function then simplifies to

$$L_{c,i}(\boldsymbol{\phi}) = \int_{-\infty}^{\infty} f(Y_{i1}, \ldots, Y_{iT}|Y_{i0}, \alpha_i; \phi)\, \mathrm{d}F(\alpha_i). \tag{4.3}$$

Conventional model fitting utilises this conditional likelihood which is equivalent to treating the lagged endogenous variable as an additional explanatory variable. The initial conditions problem occurs because the individual effects, α_i, that capture the unobserved heterogeneity are correlated with the initial state Y_{i0}. This leads to a non-standard likelihood function (4.3) which is not, in general, fully specified.

The problem can be handled by treating the Y_{i0} as endogenous and implementing the unconditional likelihood approach, which essentially models the density $f(Y_{i0}|\alpha_i)$. An important role of this treatment in devising consistent estimates in model (4.1) with no time-varying explanatory variables is fully addressed in Anderson and Hsiao (1981). The estimation of dynamic regression models with $\mathbf{X}_{it}$'s is somewhat more complicated. A change in $\mathbf{X}$ that affects the distribution of Y in the current period will continue to affect this distribution in the forthcoming period. To see this, taking backward substitution of the model (4.1) gives

$$Y_{it} = \frac{1-\gamma^t}{1-\gamma}\lambda + \gamma^t Y_{i0} + \sum_{j=0}^{t-1} \gamma^j \mathbf{X}'_{i,t-j}\boldsymbol{\beta} + \sum_{j=0}^{t-1} \gamma^j u_{i,t-j}. \tag{4.4}$$

This equation expresses the result that the current mean level of Y_{it} for each t depends directly on both past and present levels of the explanatory variables and on the initial observations Y_{i0}. Suppose now that the stochastic process which generates the Y_{it}'s has been in operation prior to the starting date of the sample. The initial state for each individual may be determined by the process generating the panel sample. Treating pre-sample history and hence the initial conditions as exogenous is questionable because the assumption $\text{cov}(Y_{i0}, \alpha_i) = 0$ implies that the individual effects, α_i, affect Y_{it} in all periods but are not brought into the model at time 0. To estimate the ML parameters correctly, we need to account for this covariance in the model.

4.4. Maximum likelihood

If we condition on the initial state at time 0, the lagged response variable can be considered as just another regressor and the standard ML approach for fitting random effects models can be used to estimate model parameters. The likelihood then consists of terms that involve the conditional distributions of Y_{it} given $Y_{i,t-1}$ and explanatory variables with correlated residuals following an error components structure. To operationalise the likelihood approach, suppose that $\mathbf{Y}_i$ is a T-element vector whose elements contain observations on all individuals at every time period after the initial period, while $\mathbf{Y}_{i,-1}$ is the corresponding vector lagged by one time period. Rewriting Equation (4.1) in vector form for each individual gives

$$\mathbf{Y}_i = \widetilde{\mathbf{X}}_i \boldsymbol{\theta} + \mathbf{u}_i, \quad i = 1, \ldots, N, \tag{4.5}$$

where $\widetilde{\mathbf{X}}_i = (\mathbf{Y}_{i,-1}\ \mathbf{e}_T\ \mathbf{X}_i), \mathbf{X}_i = (\mathbf{X}_{i1} \cdots \mathbf{X}_{iT})', \boldsymbol{\theta} = (\gamma\ \lambda\ \boldsymbol{\beta}')'$ and $\mathbf{e}_T$ is a vector of ones with order T. The covariance matrix of the combined residual term is well defined (e.g., Baltagi, 2001) and of the particular form

$$\mathbf{V}_{T \times T} = \boldsymbol{\sigma}_c^2 \bar{\mathbf{J}}_T + \sigma_\varepsilon^2 \mathbf{E}_T, \tag{4.6}$$

where $\sigma_c^2 = \sigma_\varepsilon^2 + T\sigma_\alpha^2$, $\bar{\mathbf{J}}_T = (1/T)\mathbf{e}_T\mathbf{e}_T'$ and $\mathbf{E} = \mathbf{I}_T - \bar{\mathbf{J}}_T$ are the between- and within-individual transformation matrices with order T, respectively, and $\mathbf{I}_T$ is a $T \times T$ identity matrix. The conditional log-likelihood is given by

$$\begin{aligned}\log L\big(\boldsymbol{\theta}, \sigma_\alpha^2, \sigma_\varepsilon^2\big) \propto & -\frac{N(T-1)}{2} \log\big(\sigma_\varepsilon^2\big) - \frac{1}{2} \log\big(\sigma_c^2\big) \\ & - \frac{T}{2\sigma_c^2} \sum_i \bar{u}_i^{-2} - \frac{1}{2\sigma_\varepsilon^2} \sum_i (u_{it} - \bar{u}_i)^2,\end{aligned} \tag{4.7}$$

where the $\bar{u}_i$'s are residual means over time for each i. Taking the first partial derivatives gives

$$\hat{\boldsymbol{\theta}} = (\hat{\psi} \mathbf{B}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} + \mathbf{W}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}})^{-1} (\hat{\psi} \mathbf{b}_{\tilde{\mathbf{x}}y} + \mathbf{w}_{\tilde{\mathbf{x}}y}), \tag{4.8}$$

where $\mathbf{B}$ and $\mathbf{W}$ refer to between- and within-individual variation, respectively, and $\hat{\psi} = \hat{\sigma}_\varepsilon^2 / \hat{\sigma}_c^2$. Variance components can be estimated as

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1)} \sum \sum (r_{it} - \bar{r}_i)^2, \tag{4.9a}$$

$$\hat{\sigma}_\alpha^2 = \frac{1}{N} \sum \bar{r}_i^2 - \frac{\hat{\sigma}_\varepsilon^2}{T}, \tag{4.9b}$$

where the r_{it}'s are fitted residuals for Equation (4.1).

A drawback of the ML approach is the fact that Equations (4.8) and (4.9) are the solutions of likelihood equation (4.7), but they are not necessarily the ML estimators because the solutions may lie outside the parameter space.[1] In particular, the estimate of $\hat{\sigma}_\alpha^2$ may take negative values. Applying a constraint optimisation method and using the Kuhn–Tucker conditions we can show that estimates (4.8) and (4.9) are MLEs only when $\hat{\sigma}_\alpha^2$ produces positive values. Specifically, the likelihood has the boundary solution $\hat{\sigma}_\alpha^2 = 0$ when Equation (4.9b) takes a negative value. In this case, the model fitting reduces to the estimation of model (4.1) with the usual covariance matrix $\mathbf{V} = \sigma_\varepsilon^2 \mathbf{I}_T$.

Note that σ_α^2, being a variance, cannot be negative. This implies that, even if the estimate of this variance is zero, positive between-individual variability is expected. If observed between-individual variation is very small relative to the within-individual variation, then the estimated variance of the individual effects would take value zero. This may also occur because of model misspecification as argued in Baltagi (1981) who estimates a variance component model without state dependence.

The second problem in using the conditional likelihood approach is inconsistency of parameter estimates when N is large and T is small. The inconsistency in the estimation of regression coefficients is readily derived by taking the probability limit of both sides of Equation (4.8). It is straightforward to show that

$$\operatorname*{plim}_{N\to\infty} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = (\psi^* \bar{\mathbf{B}}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} + \bar{\mathbf{W}}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}})^{-1} (\psi^* \bar{\mathbf{b}}_{\tilde{\mathbf{x}}u} + \bar{\mathbf{w}}_{\tilde{\mathbf{x}}u}), \tag{4.10}$$

where $\psi^* = \text{plim}(\hat{\psi})$, and bar notation refers to the probability limit of the corresponding variations. The second parenthesis on the right-hand side of (4.10) reduces to the vector $[p(\psi^*)\ \mathbf{0}']'$ where $p(\psi^*)$ is a positive constant given by

$$p(\psi^*) = \psi^* \varphi_T(\gamma) \sigma_{0\alpha} + \frac{1 - \varphi_T(\gamma)}{T(1-\gamma)} \sigma_c^2 (\psi^* - \psi). \tag{4.11}$$

This equation is generally non-zero, showing that parameter estimation is inconsistent. This inconsistency arises through the non-zero expectation $\sigma_{0\alpha}$ due to the initial conditions and to the inconsistent estimate of ψ reflecting the inconsistency of the variance components estimates. In fact, we can readily show that $\text{plim}(\hat{\sigma}_\varepsilon^2) > \sigma_\varepsilon^2$ and $\text{plim}(\hat{\sigma}_\alpha^2) < \sigma_\alpha^2$. Details are given in Kazemi and Davies (2002) who derive an analytical expression for the asymptotic bias and show how the bias varies with sequence lengths and with the degree of state dependence.

[1] Useful illustrations in properties of the ML estimates for a simple variance components model are presented in McCulloch and Searle (2000).

4.5. *The full likelihood*

For the full unconditional likelihood (4.2) to be useful, it is necessary to take account of the information contained in Y_{i0} for each individual. Treating the initial conditions as endogenous, it is common to suppose that the Y_{i0} are characterised by their second-order moments and by their joint moments with the individual random effects α_i; $\sigma_{0\alpha} = E(Y_{i0}\alpha_i)$, and that the expected values of the Y_{i0} differ for each individual unit i: $\mu_{i0} = E(Y_{i0})$. Although it is a plausible assumption for the μ_{i0} to be different for each i, the incidental parameters problem, associated with μ_{i0}, arises and the ML estimates are inconsistent (Anderson and Hsiao, 1982) for large values of N and small T. This will be discussed further in this paper. To derive a general unconditional likelihood function for the model from the conditional mean and variance formulation, we first readily show that

$$E(\mathbf{u}_i|y_{i0} - \mu_{i0}) = \frac{\sigma_{0\alpha}}{\sigma_0^2}(y_{i0} - \mu_{i0})\mathbf{e}_T, \tag{4.12a}$$

$$\mathrm{Var}(\mathbf{u}_i|y_{i0} - \mu_{i0}) = \sigma_u^2\bar{\mathbf{J}}_T + \sigma_\varepsilon^2\mathbf{E}_T, \tag{4.12b}$$

where σ_0^2 is the variance of Y_{i0} and

$$\sigma_u^2 = \sigma_\varepsilon^2 + T\left(\sigma_\alpha^2 - \frac{\sigma_{0\alpha}^2}{\sigma_0^2}\right). \tag{4.13}$$

Further, the covariance between the initial and subsequent error terms is given by

$$\mathrm{Cov}(\mathbf{u}_i, Y_{i0} - \mu_{i0}) = \sigma_{0\alpha}\mathbf{e}_T. \tag{4.14}$$

These show that the mean and covariance structure of various estimating methods of the random effects model (4.1), including the ML approach, conditional on the Y_{i0} and explanatory variables, are fully specified by (4.12a) and (4.12b). Supposing the start-up residuals are $u_{i0} = Y_{i0} - \mu_i$, then the covariance matrix of the vector $(u_{i0}\ u_{i1}\ \cdots\ u_{iT})$ is given by

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_0^2 & \sigma_{0\alpha}\mathbf{e}_T' \\ \sigma_{0\alpha}\mathbf{e}_T & \sigma_u^2\bar{\mathbf{J}}_T + \sigma_\varepsilon^2\mathbf{E}_T \end{bmatrix}. \tag{4.15}$$

Suppose the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ to be vectors of all coefficients involved in mean and variance structures, respectively, for the initial and subsequent state models. Substituting Equations (4.12a), (4.12b) and (4.13) into the log-likelihood

$$\log L(\boldsymbol{\theta}, \boldsymbol{\sigma}) = \sum_i \log\bigl[f(\mathbf{u}_i|u_{i0})\bigr] + \sum_i \log\bigl[f(u_{i0})\bigr], \tag{4.16}$$

and simplifying corresponding expressions, the unconditional log-likelihood can be expressed as

$$\begin{aligned}\log L(\boldsymbol{\theta}, \boldsymbol{\sigma}) \propto & -\frac{N(T-1)}{2}\log\bigl(\sigma_\varepsilon^2\bigr) - \frac{N}{2}\log\bigl(\sigma_u^2\bigr) \\ & - \frac{T}{2\sigma_u^2}\sum_i\left(\bar{u}_i - \frac{\sigma_{0\alpha}}{\sigma_0^2}u_{i0}\right)^2 - \frac{1}{2\sigma_\varepsilon^2}\sum_i\sum_t(u_{it} - \bar{u}_i)^2 \\ & - \frac{N}{2}\log\bigl(\sigma_0^2\bigr) - \frac{1}{2\sigma_0^2}\sum_i u_{i0}^2. \qquad (4.17)\end{aligned}$$

This likelihood is not useful unless we carefully model the initial conditions to find the joint distribution of Y_{i0} and α_i.

4.6. *Modelling the initial conditions as endogenous*

In this section, an overview of the literature on modelling initial conditions, together with some alternative solutions to the problem, is presented. Then an extension of the full ML approach is introduced by using a non-equality maximisation method to guarantee that the ML estimates are within the parameter space.

4.6.1. *The stationary case*

By assuming that the data-generating process can be observed in stochastic equilibrium, the history of the explanatory variables is important in the analysis of a dynamic panel data model by noting that μ_{i0} is dependent on $\mathbf{X}_{i0}, \mathbf{X}_{i,-1}, \mathbf{X}_{i,-2}, \ldots$. It may be seen from Equation (4.12a) that the dependence is not easily removed by conditioning on $Y_{i0} - \mu_{i0}$ as in ordinary time-series models because the conditional likelihood still depends on μ_{i0}. New assumptions then have to be made about the effects of the unobserved past history of the explanatory variables in order to employ the ML approach. This may not always yield realistic results, especially when the observed series is short. But these assumptions are crucial for obtaining correct parameter estimates. If the assumptions about the way that the $\mathbf{X}$'s are generated are incorrect, then consistent results are not guaranteed.

Consider a typical case that in which the stochastic process generating Y_{it} has been in operation for a sufficiently long time period in the past before the process is observed. Suppose the distribution of the Y_{i0}'s depends upon the process and assume that the panel data follow an initial stationary process. From Equation (4.4), the start-up observations can be modelled as a function of individual random effects, α_i, the present, $\mathbf{X}_{i0}$, and the unobserved past of the explanatory variables $\mathbf{X}_{i,-1},\ldots,$ and a serially uncorrelated disturbance ε_{i0}:

$$Y_{i0}=\frac{\lambda}{1-\gamma}+\mathbf{X}_{i0}^{*\prime}\boldsymbol{\beta}+\frac{\alpha_i}{1-\gamma}+\varepsilon_{i0}, \tag{4.18}$$

where $\mathbf{X}_{i0}^{*}=\mathbf{X}_{i0}/(1-\gamma L)$, L is the lag operator, and $\varepsilon_{i0}=\sum_{j=0}^{\infty}\gamma^{j}\varepsilon_{i,-j}$. Assuming the process is in equilibrium, the distribution of Y_{i0} depends on the distribution of α_i, ε_{it}, and $\mathbf{X}_{it}$ for all i and t. More specifically, the expectation μ_{i0} depends on the pre-sample mean values of these $\mathbf{X}_{it}$'s. One suggestion is to treat the cumulative effect of past $\mathbf{X}$'s as an unknown parameter (Anderson and Hsiao, 1982) and let the means μ_{i0} be the first two terms of the right-hand side on Equation (4.18); that is

$$Y_{i0}=\mu_{i0}+\frac{\alpha_i}{1-\gamma}+\varepsilon_{i0}. \tag{4.19}$$

Then the ML estimates of the vector parameter $(\boldsymbol{\theta},\mu_{10},\ldots,\mu_{N0},\sigma_{\alpha}^{2},\sigma_{\varepsilon}^{2})$ can be obtained from the unconditional likelihood (4.17) with

$$\begin{aligned}\sigma_0^2&=\frac{\sigma_\alpha^2}{(1-\gamma)^2}+\frac{\sigma_\varepsilon^2}{(1-\gamma^2)},\\ \sigma_{0\alpha}&=\frac{\sigma_\alpha^2}{1-\gamma}.\end{aligned} \tag{4.20}$$

This treatment of the means of Y_{i0}, however, leads to inconsistency in the parameter estimates because of the problem of incidental parameters. To overcome this problem, making some assumptions about the generation of the $\mathbf{X}_{it}$'s in the pre-sample period is required. This is essentially the case considered by Bhargava and Sargan (1983), Maddala (1987), and Ridder and Wansbeek (1990).

Bhargava and Sargan (1983) assume that the Y_{i0}'s are generated by the same data-generating process as that which generates the subsequent panel data (see also Hsiao *et al.*, 2002). By assuming that the panel data follow an initial stationary process, the suggested model for the initial state is given by

$$Y_{i0}=\lambda_0+\sum_{t=0}^{T}\mathbf{X}_{it}'\boldsymbol{\eta}_t+u_{i0}, \tag{4.21}$$

where

$$u_{i0}=\upsilon_{i0}+\frac{\alpha_i}{1-\gamma}+\varepsilon_{i0},\quad i=1,\ldots,N, \tag{4.22}$$

and $\upsilon_{i0} \sim$ i.i.d.$(0, \sigma_v^2)$. The variance of u_{i0} and its covariance with u_{it} are

$$\begin{aligned} \operatorname{var}(u_{i0}) &= \sigma_v^2 + \frac{\sigma_\alpha^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{(1-\gamma^2)}, \\ \operatorname{cov}(u_{i0}, u_{it}) &= \frac{\sigma_\alpha^2}{1-\gamma}, \quad t = 1, \ldots, T. \end{aligned} \tag{4.23}$$

Then the unconditional likelihood function can be derived from the joint distribution of the vector residuals $(u_{i0}, u_{i1}, \ldots, u_{iT})$.

Equation (4.21), however, is an econometric construction to try to generate the Y_{i0} using an optimal linear predictor, suggested by Chamberlain (1984), of Y_{i0} conditional on all explanatory variables $\mathbf{X}_{i0}, \ldots, \mathbf{X}_{iT}$. This treatment is applicable if the pre-sample $\mathbf{X}_{it}$ have linear conditional expectations. If we assume the random process starts in period $t = 0$ and the $\mathbf{X}_{it}$ are weakly exogenous, then these lead us to have specific reduced forms of

$$Y_{i0} = \lambda_0 + \mathbf{X}_{i0}'\boldsymbol{\beta}_0 + \upsilon_{i0} + \frac{\alpha_i}{1-\gamma} + \varepsilon_{i0}, \quad \text{for } i = 1, \ldots, N, \tag{4.24}$$

where $\mathbf{X}_{i0}$ are explanatory variables of the start-up process containing information on the period $t = 0$.

A modified likelihood function for the initial conditions equation of the first state is suggested in Maddala (1987). To derive an expression for $\mathbf{X}_{i0}^*$, assuming a stationary normal process for $\mathbf{X}_{it}$, implies that $\mathbf{X}_{i0}^*$ can be decomposed into independent components $\sum_{t=1}^{T} \mathbf{X}_{it}'\boldsymbol{\pi}_t$ and an error term. Inserting this in Equation (4.18) and substituting Y_{i0} in (4.1) for $t = 1$, the process is now modified for the first state as the initial conditions equation

$$Y_{i1} = \frac{\lambda}{1-\gamma} + \mathbf{X}_{i1}'\boldsymbol{\beta} + \sum_{t=1}^{T} \mathbf{X}_{it}'\boldsymbol{\delta}_t + u_{i1}^*, \tag{4.25}$$

where $u_{i1}^* = \frac{\alpha_i}{1-\gamma} + \varepsilon_{i1}$, and $\varepsilon_{i1} \sim$ i.i.d.$(0, \sigma_*^2)$. The ML estimation now proceeds with Equation (4.25) for Y_{i1} and (4.1) for $t > 1$ with identifiable parameters $\lambda, \gamma, \beta, \boldsymbol{\delta}_t, \sigma_\alpha^2, \sigma_\varepsilon^2$ and σ_*^2, based upon the joint distribution of $(u_{i1}^*, u_{i2}, \ldots, u_{iT})$. Maddala's approach comes to much the same result as that outlined by Lee (1981) with a little change in Equation (4.25).

Nerlove and Balestra (1996) obtain the unconditional likelihood function for the vector $(Y_{i0}, Y_{i1}, \ldots, Y_{iT})$ by assuming that the dynamic relationship is stationary. They first take deviations from individual means to eliminate the intercept term. Then, in the model (4.1) with only one explanatory variable, they assume the process which generates X_{it} follows a stationary time series model for all individuals i. Further, they assume that the Y_{i0} are normally distributed with zero means and common variance

$$\sigma_0^2 = \frac{\beta^2\sigma_x^2}{1-\gamma^2} + \frac{\sigma_\alpha^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{1-\gamma^2}. \tag{4.26}$$

The probability density function of the initial observations then enters the final term in the conditional likelihood (4.7) to give the unconditional log-likelihood[2]

$$\begin{aligned}\log L\bigl(\gamma, \beta, \sigma_\alpha^2, \sigma_\varepsilon^2\bigr) \propto & -\frac{N(T-1)}{2}\log\bigl(\sigma_\varepsilon^2\bigr) - \frac{1}{2}\log\bigl(\sigma_c^2\bigr) \\ & - \frac{T}{2\sigma_c^2}\sum_i \bar{u}_i^2 - \frac{1}{2\sigma_\varepsilon^2}\sum_i (u_{it} - \bar{u}_i)^2 \\ & - \frac{N}{2}\log\bigl(\sigma_0^2\bigr) - \frac{1}{2\sigma_0^2}\sum_i Y_{i0}^2. \end{aligned} \tag{4.27}$$

The term $\sigma_{0\alpha}$ does not appear in likelihood (4.27) which shows that the initial observations are independent of the random effects. If Y_{i0} is dependent on α_i, estimation of the parameters of all subsequent Y's involves this dependency. This dependency term is not easily removed because the conditional distribution $f(Y_{i1} \cdots Y_{iT}|Y_{i0})$ in full likelihood (4.2) depends on the joint distribution of Y_{i0} and α_i.

A potential problem when modelling the Y_{i0} suggested in the literature is that the corresponding initial equations are derived by making the assumption that the Y_{i0} are drawn from a stationary distribution. This assumption is not necessarily appropriate in many cases in practice and may not always yield realistic results. A special case may arise when at least one shock comes shortly before the sample starts, jolting the data generating process out of stationarity. Furthermore, in many panel data applications, the time series components have strongly evident non-stationarity, a feature which has received little attention in the literature of panel data analysis with initial conditions. In such a process, the assumption of stationarity of the process is unattractive, in particular when explanatory variables drive the stochastic process. For these reasons we propose a pragmatic solution as follows.

[2] Nerlove and Balestra (1996) use a special transformation on variables to derive the likelihood function. It can readily be shown that the likelihood with transformed data is the same as the likelihood (4.27).

4.6.2. A pragmatic solution

A realistic way of dealing with the problem of initial conditions is to add flexible reduced form equations for the initial period similar to the dynamic equations, but without the lagged response variables. The coefficients are allowed to be different from the coefficients in the dynamic equations, and the error terms are allowed to have a different covariance structure. This is the straightforward generalisation of the solution suggested by Heckman (1981) who argues that the approach performs with greater flexibility than other random effects models.

More specifically, suppose that the Y_{i0}'s are the start-up values of the stochastic process and that the $\mathbf{X}_{i0}$'s are corresponding explanatory variables consisting of information on the period $t = 0$. Depending on different specifications of the joint distribution of random variables Y_{i0} and α_i, we treat the initial observations as endogenous and impose a reduced equation form which describes the effects of the variables $\mathbf{X}_{i0}$'s on these observations. It becomes necessary, therefore, to set up a distinct start-up regression model

$$Y_{i0} = \lambda_0 + \mathbf{X}'_{i0}\boldsymbol{\beta}_0 + u_{i0}, \quad i = 1, \ldots, N, \tag{4.28}$$

where the $\mathbf{X}_{i0}$'s are supposed to be uncorrelated with u_{i0}, the initial start-up error terms u_{i0} being distributed randomly having zero means and a common variance, and the u_{i0} and ε_{it} to be uncorrelated for all $t > 0$. The distribution of the initial values Y_{i0} are then correctly specified with means

$$\mu_{i0} = E(Y_{i0}|\mathbf{X}_{i0}) = \lambda_0 + \mathbf{X}'_{i0}\boldsymbol{\beta}_0, \tag{4.29}$$

and a common variance $\sigma_0^2 = \mathrm{var}(Y_{i0}|\mathbf{X}_{i0})$. We suppose the covariance between the initial error and random effects to be a non-zero constant for all i given by $\sigma_{0\alpha}$. With these assumptions, Equation (4.29) shows that, as time goes on, the impact of start-up errors u_{i0} affects all forthcoming values of Y_{it} through the non-zero covariance $\sigma_{0\alpha}$. In other words, α_i affects Y_{it} in all subsequent periods, including Y_{i0}. We note that there is no restriction on the components of the covariance matrix $\boldsymbol{\Omega}$.

Supposing $\widetilde{\mathbf{X}}_{i0} = (1\ \mathbf{X}'_{i0})$ and rewriting Equation (4.28) in vector form gives

$$Y_{i0} = \widetilde{\mathbf{X}}_{i0}\boldsymbol{\theta}_0 + u_{i0}, \quad i = 1, \ldots, N, \tag{4.30}$$

where $\boldsymbol{\theta}_0 = (\lambda_0\ \boldsymbol{\beta}'_0)'$. It follows from the unconditional log-likelihood (4.17) and taking the first partial derivatives with respect to the vector parameter $(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \sigma_0^2, \sigma_{0\alpha}, \sigma_\alpha^2, \sigma_\varepsilon^2)$ that the solutions of the likelihood equations are given by

$$\hat{\boldsymbol{\theta}}_0 = \left(\frac{1}{N}\sum_i \widetilde{\mathbf{X}}_{i0}\mathbf{X}'_{i0}\right)^{-1} \frac{1}{N}\sum_i \widetilde{\mathbf{X}}_{i0}\left(Y_{i0} - \frac{T\hat{\sigma}_{0\alpha}}{\hat{\sigma}_\varepsilon^2 + T\hat{\sigma}_\alpha^2}\bar{r}_i\right), \tag{4.31a}$$

$$\hat{\boldsymbol{\theta}} = (\hat{\psi}\mathbf{B}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} + \mathbf{W}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}})^{-1}\left(\hat{\psi}\mathbf{b}_{\tilde{\mathbf{x}}y} + \mathbf{w}_{\tilde{\mathbf{x}}y} - \hat{\psi}\frac{\hat{\sigma}_{0\alpha}}{\hat{\sigma}_0^2}\frac{1}{N}\sum_i \bar{\tilde{\bar{\mathbf{X}}}}_i r_{i0}\right), \tag{4.31b}$$

where $\hat{\psi} = \hat{\sigma}_\varepsilon^2/\hat{\sigma}_u^2$, the r_{i0}'s are fitted residuals corresponding to the start-up regression equation (4.30), the r_{it}'s are fitted residuals for Equation (4.1) and $\bar{r}_i$'s are their individual means over time. We can readily show that the parameter estimates of the variances and the covariance between Y_{i0} and α_i are

$$\hat{\sigma}_0^2 = \frac{1}{N}\sum_i r_{i0}^2, \tag{4.32a}$$

$$\hat{\sigma}_{0\alpha} = \frac{1}{N}\sum_i r_{i0}\bar{r}_i, \tag{4.32b}$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1)}\sum_i (r_{it} - \bar{r}_i)^2, \tag{4.32c}$$

$$\hat{\sigma}_u^2 = \frac{T}{N}\sum_i \left(\bar{r}_i - \frac{\sigma_{0\alpha}}{\sigma_0^2}r_{i0}\right)^2. \tag{4.32d}$$

The estimate $\hat{\sigma}_\alpha^2$ can then be derived from Equation (4.13). We note that the conditional ML estimates are a special case of these equations where $\sigma_{0\alpha}$ is assumed to be zero. More specifically, these equations for $\hat{\boldsymbol{\theta}}$, $\hat{\sigma}_\varepsilon^2$, and $\hat{\sigma}_u^2$ reduce to (4.8), (4.9a) and (4.9b) if the initial conditions are generated exogenously in the model. Besides, the parameters of the initial start-up regression (4.30) can be estimated from the ML estimates using the initial observations only.

It follows from Equation (4.4) that the stochastic variable $\sum_i \bar{\tilde{\bar{\mathbf{X}}}}_i r_{i0}/N$ tends to the non-zero vector $(p\ \mathbf{0}')'$ in probability for large N, where p is a positive constant given by

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_i \overline{Y}_{i,-1}u_{i0} = \varphi_T(\gamma)\sigma_0^2 + \frac{1-\varphi_T(\gamma)}{1-\gamma}\sigma_{0\alpha}, \tag{4.33}$$

for fixed T. From Equations (4.11) and (4.33), we can readily demonstrate the consistency of parameter estimates.

Further, by a similar argument to that in Section 4.4, we can show that the solutions of likelihood equations (4.31) and (4.32) are not the ML estimates because these solutions may lie outside the parameter space. Specifically, the issue arises with the restrictive bound $0 < \psi \leq 1$ which ensures that the estimates are within the parameter space. But the estimates are not MLEs when the estimate $\hat{\psi}$ is out of this range. Using an inequality constraint for solving maximisation problems and Kuhn–Tucker's conditions, it can be shown that Equations (4.31) and (4.32) are the ML estimates, provided that the estimate of $\hat{\psi}$ lies within the interval $(0, 1)$. In the case $\psi \geq 1$, we can show that the only active constraint is the optimal point $(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \sigma_\varepsilon^2, \sigma_0^2, \sigma_{0\alpha})$ that lies on the boundary $\psi = g(\sigma) = 1$ of the corresponding constraint. This means that the inequality maximisation problem reduces to the method of Lagrangian multipliers in the familiar setting of equality constraints. On the boundary $\psi = 1$, consistent ML estimates are reduced forms of Equations (4.31) and (4.32) when $\hat{\psi}$ equals one. Further, for the variance component estimates, we derive the same equations for σ_0^2 and $\sigma_{0\alpha}$ while the σ_ε^2 can be consistently estimated by

$$\sigma_\varepsilon^2 = \frac{1}{NT} \sum_i \sum_t \left(u_{it} - \frac{\sigma_{0\alpha}}{\sigma_0^2} u_{i0} \right)^2 . \tag{4.34a}$$

The variance of the individual effects can then be estimated by

$$\sigma_\alpha^2 = \frac{\sigma_{0\alpha}^2}{\sigma_0^2} . \tag{4.34b}$$

Equation (4.34b) shows that the maximisation of the unconditional likelihood (4.17) with respect to the parameters that are within the parameter space results in non-negative estimates of variance, if there exists a positive correlation between the initial outcomes and the random effects.

4.7. *Empirical analysis*

This section considers an empirical analysis to examine the consequences of ignoring initial conditions when fitting the dynamic regression (4.1), followed by a variety of different models for the Y_{i0} in the presence of random effects. Specifically, we re-examine estimates from Nerlove's (2002) study where he presents the results of the unconditional ML approach and the standard estimates conditional on initial outcomes on the finding of economic growth rate convergence. The purpose of this example is to show that the unconditional ML approach does not always give reliable results and to illustrate a way of improving it.

Now, we follow the literature and fit Ordinary Least Squares (OLS) regression, fixed effects, and the following models to the data using the ML procedure to illustrate the important role of the initial conditions empirically:

- M1: The unconditional ML of Nerlove's method if $\sigma_{0\alpha} \neq 0$.

- M2: The initial conditions are exogenous in the unrestricted version of $\boldsymbol{\Omega}$. That is, we use Equation (4.29) and assume the covariance $\sigma_{0\alpha}$ to be zero.

The initial conditions are endogenous and the X_{it} are weakly exogenous:

- M3: the pragmatic approach.
- M4: the initial stationarity of the process; i.e. using Equation (4.24).

The initial conditions are endogenous and the X_{it} are strictly exogenous:

- M5: Bhargava and Sargan's (1983) approach.
- M6: an alternative approach to M5 which is explained below.

The specification of most of these models is extensively explained in the previous sections. An additional model is an alternative to Bhargava and Sargan's (1983) approach, given by M6, which assumes the reduced form approximation (4.21) for Y_{i0} and allows the residuals u_{i0} to be unrestrictedly identified with common variance σ_0^2 and the covariance $\sigma_{0\alpha}$.

After fitting the parameters of these candidate models, an important empirical question is clearly how to select an appropriate model for the analysis of each data set. In particular, we use Akaike's Information Criterion (*AIC*) and Bayesian Information Criterion (*BIC*) to select the best fitted model.

4.7.1. Dynamic growth panel data models

This section considers recent empirical work on classical convergence which refers to an empirical specification of the Solow–Swan neoclassical growth model. Many empirical growth studies into cross-country convergence follow Mankiw *et al.* (1992) in using a log-linear approximation to the growth model to specify regression equations and to investigate the question of convergence of countries to the steady state. Within a large number of these studies, considerable attention has been paid to properties of the various parameter estimates in fixed effects specifications of regression models. It is argued that the incorrect treatment of country-specific effects, representing technological differences between countries, gives rise to omitted variable bias. This implies that the various parameter estimates of fixed effects models for dynamic panel data are inconsistent (Nickell, 1981; Kiviet, 1995), since the omitted variable is correlated with the initial level of income. In the estimation of empirical growth models with random effects, Nerlove (2000) highlights the issue of small sample bias by assuming that the stochastic process that generates Y_{it} is stationary and that the initial per capita *GDP* are generated by the process suggested in Nerlove and Balestra (1996).

The application of dynamic panel data to the estimation of empirical growth models was originally considered by Knight *et al.* (1993) and Islam (1995). In these contributions, the growth convergence equation is derived from assumptions about a production function and inclusion in the specification of the savings rate, s, the population growth rate, n, the rate of technical progress, g, and the depreciation rate, δ.[3] Suppose Y_{it} to be the logarithm of per capita *GDP* for country i at time t, then the dynamic growth model for panel data is given by

$$Y_{it} = \lambda + \gamma Y_{i,t-1} + \beta_1 \log(s_{it}) + \beta_2 \log(n_{it} + g + \delta) + \alpha_i + \varepsilon_{it}, \tag{4.35}$$

where α_i is a country-specific effect and ε_{it} is the time-varying error term. The coefficient of the lagged per capita *GDP* is a function of the rate of convergence r: namely, $\gamma = \exp(-r\tau)$, where τ is the time interval. If the parameter γ is estimated positively much less than one, then the results support the growth convergence hypothesis; i.e. countries with low initial levels of real per capita income are growing faster than those with high initial values. Whereas, if this coefficient is close to one or slightly larger than one, then the initial values have little or no effect on subsequent growth.

In empirical investigations of the rate of economic growth convergence, a simple restricted form of dynamic model (4.35) may be considered. This form comes from the constant returns to scale assumption in the Solow–Swan model implying that β_1 and β_2 are equal in magnitude and opposite in sign. Equation (4.35) then reduces to

$$Y_{it} = \lambda + \gamma Y_{i,t-1} + \beta X_{it} + \alpha_i + \varepsilon_{it}, \tag{4.36}$$

where $X_{it} = \log(s_{it}) - \log(n_{it} + g + \delta)$.

The basic data are annual observations for 94 countries over the period 1960–85 taken from the Penn World Tables in Summers and Heston (1991). In the empirical application of the growth regression model (4.35), to avoid modelling cyclical dynamics, most growth applications consider only a small number of time periods, based on annual averages. Working with regular non-overlapping intervals of five years, the cross-sections correspond to the years t = 1965, 1970, 1975, 1980, and 1985. For the variable Y_{it}, the observations of each cross-section correspond exactly to the year t, while s_{it} and n_{it} correspond to averages over the period $t - 1$

[3] The derivation of the growth convergence equations is available in many recent papers and hence is not reproduced here.

Table 4.1. Parameter estimates with the OLS regression, fixed and random effects

Parameter	OLS	Fixed effects	Conditional ML	Unconditional ML*
γ	0.9487 (0.0090)	0.7204 (0.0237)	0.93390.0123	0.9385 (0.0105)
β	0.1244 (0.0108)	0.1656 (0.0193)	0.13700.0132	0.1334 (0.0124)
ρ	–	–	0.1133 (0.0497)	0.1288 (0.0456)
σ^2	–	–	0.0194 (0.0013)	0.0197 (0.0013)
Obs.	470	470	470	564

The intra-class correlation $\rho = \sigma_\alpha^2/\sigma^2$ where $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$.
Standard errors are in parentheses.
*The estimates are reported by Nerlove (2000).

to t. As in the empirical growth models, it is assumed that g and δ, summing to 0.05, are constant across countries.[4]

Before proceeding to the random effects specification, we take a look at the fixed effects that explicitly model a random process in the residuals by applying the specific transformation of data from their averages over a time period. We then estimate the empirical model (4.36) with *OLS* regression, and estimate the random effects using likelihood (4.7) which assumes the Y_{i0}'s are exogenous. The results are shown in Table 4.1, together with those obtained by Nerlove (2000) using an unconditional version of the likelihood which includes the 1960 *GDP* per capita.

The estimates of the lagged *GDP* coefficient γ are statistically significant resulting in strong evidence for state dependence. The estimates $\hat{\gamma}$ are quite large in the random effects specification relative to the fixed effects. The estimate of the intra-class correlation ρ, representing the random effects specification, is relatively large with respect to its standard error, suggesting that an *OLS* analysis of these data would be likely to yield misleading results. In fact, if the unconditional ML estimates of parameters are near their true values, the *OLS* overestimates the true γ while the fixed effects underestimate it. This implies that other estimates of γ, presented below for different methods, may fall between these two estimates.

Our main emphasis for the analysis of economic growth data is to select, from a set of candidate models, an appropriate approximating model

[4] The standard assumption in the literature is that $g = 0.02$ and $\delta = 0.03$, but researchers report that estimation results are not very sensitive to changes in these values (Caselli *et al.*, 1996). An alternative procedure is to estimate these coefficients inside the growth equation (Lee and Pesaran, 1998).

Table 4.2. Parameter estimates by using an unconditional ML of Nerlove's method in the case of non-zero covariance $\sigma_{0\alpha}$

Parameter	M1
γ	0.8774 (0.0302)
β	0.1464 (0.0107)
λ	0.9119 (0.2255)
σ_{ε}^2	0.0200 (0.0014)
σ_{α}^2	0.0088 (0.0048)
Obs.	564

Standard errors are in parentheses.

that best fits the empirical data set. Specifically, six candidate models, explained earlier in this section, can be fitted by ML methods and ranked by the use of two popular criteria, *AIC* and *BIC*. The first model which requires a reconsideration is the unconditional ML of Nerlove's method.

An important issue in modelling initial conditions is that when the statistical relationship between the random effects and initial conditions is not correctly specified, estimation results will be suspected. In fact, we know from Sections 4.4 and 4.6 that the conditional ML estimate of γ is seriously upwardly biased for small sequence lengths, and that the unconditional ML $\hat{\gamma}$ is an unbiased estimate of γ. For these strong theoretical reasons, the estimate of γ for unconditional ML would be expected to be smaller than the conditional estimate, which is not surprisingly true in this analysis. This implies that we are confronted with a specification that is not unbiased. This critical point has not been addressed by Nerlove (2000, 2002). Particularly, the large value of $\hat{\gamma}$, representing the expected positive bias in the unconditional ML estimate, reveals that it is inappropriate both to assume that the covariance σ_{0a}, which equals $\sigma_{\alpha}^2/(1-\gamma)$, is zero and to ignore it when deriving the unconditional likelihood. To improve the poor performance of the estimation method, it would seem more realistic to involve this covariance in the likelihood function. Table 4.2 shows the results from re-estimating convergence equation (4.36) with this modification in the unconditional likelihood.

As can easily be seen, the parameter estimates are substantially changed. As we expected, the estimate $\hat{\gamma}$ is lower than the conditional estimate, shown in Table 4.1, suggesting an improvement in results which is clearly consistent with the theoretical analysis that the unconditional ML estimate is upwardly biased. While the coefficient of the lagged *GDP* seems to show the most severe bias, the estimate of β is also biased.

One important point about this modified version refers to the dependency term $\sigma_{0\alpha}$ representing the association between the initial observations and the subsequent process. This coefficient is given by $\sigma_\alpha^2/(1-\gamma)$ and can now be estimated consistently from the results of Table 4.2 with the value 0.0716 and its standard error 0.0233. This shows a significant positive association and therefore suggests including the dependency term $\sigma_{0\alpha}$ in the unconditional likelihood to estimate model parameters.

Although the above approach is undoubtedly an improvement on the estimation method that ignores the covariance $\sigma_{0\alpha}$ in the likelihood, it still requires further consideration. A crucial issue in modelling initial conditions, to ensure more accurate and consistent results, relates to the correct specification of the mean values of Y_{i0}, which represent the relationship between the initial observations and the explanatory variable X_{it}. It should be noted at this point that the preceding analysis does not examine the structure of the mean values of Y_{i0} and simply ignores the effects of μ_{i0} in the data analysis. To examine the effects of this shortcoming, we specifically consider the two approaches proposed in Section 4.6. The results of the maximisation of the likelihood for the economic growth model (4.36), with both initial stationarity of the process and unrestricted versions of the covariance matrix $\boldsymbol{\Omega}$, are in Table 4.3. In fitting the models for the unrestricted case, the parameter ψ takes a value in the eligible interval $(0, 1]$ which shows that we can estimate the parameters using estimating equations (4.31) and (4.32). In the initial stationarity case, the estimates for parameters σ_a^2 and γ are constrained to be, respectively, non-negative and restricted to interval $(0, 1)$ during the numerical maximisation process.[5]

Assuming the X_{it}'s are weakly exogenous and involve only X_{i0} in the initial model, the results of a special case when the initial observations are independent of the random effects are given in the first column of Table 4.3 (model M2). Here, the parameters of the subsequent model (4.36) and the initial start-up regression (4.28) can be estimated, respectively, conditionally on Y_{i0} by likelihood (4.7) and from the ML estimates using the initial observations only. When Y_{i0} is assumed to be in stationary equilibrium (model M4), the estimated state dependence parameter $\hat{\gamma}$ is almost unchanged in comparison with the pragmatic approach (model M3) whereas the estimate $\hat{\beta}$, in contrast, falls by about 12%. This suggests that if the expected values of Y_{i0} are assumed inappropriately in reduced forms, then

[5] The likelihood function may not have a unique maximum within the parameter space. To avoid maximisation routines converging to a local maximum, we run the program with various starting values for the unknown parameters. If different starting values converge to the same point, we confidently choose that point. Otherwise, the points with the higher likelihood are the points of interest.

Table 4.3. Estimation results for different models when X_{it}'s are weakly exogenous

Parameter	M2	M3	M4
Parameter estimates for the initial conditions equation			
β_0	0.6987 (0.0789)	0.6424 (0.0757)	0.6034 (0.0823)
λ_0	7.0635 (0.0757)	7.0872 (0.0753)	7.2578 (0.0798)*
Parameter estimates for the subsequent panel data equation			
γ	0.9339 (0.0122)	0.8393 (0.0207)	0.8397 (0.0194)
β	0.1370 (0.0131)	0.1957 (0.0147)	0.2217 (0.0141)
λ	0.5162 (0.0874)	1.2002 (0.1523)	1.1631 (0.1438)
Variances and covariances estimates			
σ_0^2	0.4353 (0.0635)	0.4377 (0.0643)	0.5111 (0.0732)*
$\sigma_{0\alpha}$	–	0.0576 (0.0134)	0.0546 (0.0128)*
σ_ε^2	0.0172 (0.0013)	0.0153 (0.0011)	0.0160 (0.0012)
σ_α^2	0.0022 (0.0010)	0.0103 (0.0034)	0.0087 (0.0029)
σ_υ^2	–	–	0.1163 (0.0422)
Obs.	564	564	564

Standard errors are in parentheses.

*The constant term is $\lambda/(1-\gamma)$. Two parameters $\sigma_{0\alpha}$ and σ_0^2 are estimated by Equation (4.23). Standard errors are constructed via the delta method.

the ML approach can lead to inconsistent results, especially in the estimation of the coefficient on X_{it}.

Table 4.4 summarises the results of estimating equation (4.36) when the X_{it}'s are assumed to be strictly exogenous. The estimation results using the approach of Bhargava and Sargan (1983), presented in the first column, together with the alternative approach to M5 (the second column) show that inclusion of the explanatory variables for each period in the initial model does not appreciably change the estimated model compared to using only X_{i0} in the initial model. Indeed, the estimate of β actually decreased slightly. By adding more regressors in the reduced model for Y_{i0}, as we would expect, the variance estimate σ_0^2 representing the variation in Y_{i0} is rather decreased. But there is no improvement in the estimation of covariance $\sigma_{0\alpha}$.

In summary, the results from various assumptions on Y_{i0} support the growth convergence hypothesis, conditional on savings and population growth rates, but illustrate the rather different estimates of the rates of convergence. The estimates $\hat{\gamma}$ are less than one for fitted models, suggesting that the countries with low initial *GDP* per capita values are growing

Table 4.4. Estimation results for different models when X_{it}'s are strictly exogenous

Parameter	M5	M6
Parameter estimates for the initial conditions equation		
β_0	0.8585 (0.2137)	0.7071 (0.1843)
β_1	0.0670 (0.3138)	−0.1771 (0.2737)
β_2	−0.2802 (0.2757)	0.0206 (0.2413)
β_3	−0.0887 (0.1260)	−0.0892 (0.1077)
β_4	−0.2074 (0.2254)	0.2909 (0.2074)
β_5	0.1805 (0.1909)	−0.0730 (0.1716)
λ_0	7.3394 (0.0851)*	7.0386 (0.0908)
Parameter estimates for the subsequent panel data equation		
γ	0.8384 (0.0202)	0.8390 (0.0206)
β	0.2082 (0.0145)	0.1976 (0.0155)
λ	1.1863 (0.1504)	1.2017 (0.1517)
Variances and covariances estimates		
σ_0^2	0.4776 (0.0751)*	0.4280 (0.0653)
$\sigma_{0\alpha}$	0.0579 (0.0142)*	0.0572 (0.0135)
σ_ε^2	0.0156 (0.0012)	0.0153 (0.0012)
σ_α^2	0.0094 (0.0032)	0.0103 (0.0034)
σ_υ^2	0.1084 (0.0392)	–
Obs.	564	564

Standard errors are in parentheses.

*The constant term is $\lambda/(1-\gamma)$. Two parameters $\sigma_{0\alpha}$ and σ_0^2 are estimated by Equation (4.23). Standard errors are constructed via the delta method.

faster than those with high values. As the speed of convergence is inversely proportional to the relative size of $\hat{\gamma}$, the conditional ML estimate leads to a downwardly biased estimate of this rate. The unconditional estimates also give rise to inconsistent results unless the statistical relationships between the random effects, initial conditions, and explanatory variables are correctly specified. It specifically requires having both $\sigma_{0\alpha}$ and μ_{i0} in the likelihood functions, unlike most previous approaches.

4.7.2. Model selection

The model selection problem is to select, from a candidate set of models, the one that best fits the data set based on certain criteria. The substantial advantages in using the following criteria, are that they are valid for non-

Table 4.5.* *Comparison of models for dynamic growth panel data

Model	M	−2 log(L)	*AIC*	Δ_k	*BIC*
M1	5	−235.7739	−225.7739	139.532	−204.0986
M2	8	−341.278	−325.278	40.0279	−290.5976
M3	9	−383.3059	−365.3059	0	−326.2904
M4	7	−345.8728	−331.8728	33.4331	−301.5274
M5	12	−353.6404	−329.6404	35.6655	−277.6197
M6	14	−358.3761	−330.3761	34.9298	−269.6853

nested models, and that the ranking of models using them also helps to clarify the importance of model fitting.

The most common measure for choosing among different competing models for a given data set is *AIC* defined for the *k*th model as

$$AIC_k = -2\log\big\{L(\hat{\boldsymbol{\phi}})\big\}_k + 2M, \quad k = 1, \ldots, K, \tag{4.37}$$

where M is the number of model parameters and $\hat{\boldsymbol{\phi}}$ is the ML estimate of model parameters. When selecting among K competing models, it seems reasonable to say that the larger the maximum log-likelihood, the better the model, or, the model that yields the smallest value of AIC_k is the preferred model. If none of the models is good, *AIC* attempts to select the best approximating model of those in the candidate set. Thus, it is extremely important to ensure that the set of candidate models is well-substantiated. Because the *AIC* value is on a relative scale, the *AIC* differences

$$\Delta_k = AIC_k - \min AIC_k, \tag{4.38}$$

are often reported rather than the actual values. This simple rescaling to a minimum relative *AIC* value of zero makes comparisons between the best fitting model and other candidate models easy. A larger Δ_k reflects a greater distance between models. Another criterion likely to be is *BIC* defined as

$$BIC_k = -2\log\big\{L(\hat{\boldsymbol{\phi}})\big\}_k + M\log(n), \quad k = 1, \ldots, K, \tag{4.39}$$

where n is the total number of observations. The *BIC* automatically becomes more severe as sample size increases, but provides similar results to the *AIC* for small n.

The values of *AIC* and *BIC* together with the number of estimable parameters (M) for different fitted models of the growth data, are presented in Table 4.5.

The results of Table 4.5 show that two models, M3 and M4, are more appropriate, given the data, while M1 and M2 are unlikely to be preferred.[6] Specifically, the Δ values indicate that the model with the minimum *AIC* value of the six models is the pragmatic model. Most importantly, comparing M2 and M3 suggests that it is questionable to assume $\sigma_{0\alpha} = 0$. It is easily seen that estimated models which take the X_{i0} into account are reasonable models in comparison with Ml, which does not. In general, the less restrictive reduced form model provides a significantly better empirical fit to the data. It is interesting to note that the two criteria *AIC* and *BIC* produce slightly different results, although in both cases M3 has the smallest value while M6 the largest.

Summarising the empirical results reveals that there is strong evidence of positive state dependence with various assumptions on the data-generating processes. Ignoring the endogeneity of Y_{i0} results in upward bias of the state dependence and a downward bias in the coefficients of explanatory variables. The interpretation of the empirical models, based only on the exogeneity assumed for the initial conditions, may be misleading. By fitting various models for Y_{i0} and comparing the results, we conclude that there is not only theoretical but also empirical evidence to suggest that the initial conditions problem plays a crucial role in obtaining more reliable results.

4.8. Recommendations

Although the impact of initial conditions on subsequent outcomes of a dynamic process is widely discussed in the context of state-dependent models, it is not fully understood in statistical modelling. A basic problem in fitting these models with random effects, as is well known, is that the standard likelihood estimates can be substantially biased at least asymptotically. To avoid this, the model can be extended by adding a set of flexible equations for the initial outcome. The ML approach may then help in devising consistent results for model parameters if the joint likelihood of initial errors and residual heterogeneity in a subsequent sequence of panel data is correctly specified. Specifically, there is a need for the correct specification of the relationship between the individual-specific effects, initial conditions and explanatory variables.

[6] When the *AIC* values are negative, as they are here, higher numbers in absolute values are preferred. For more detail see Burnham and Anderson (1998).

These concepts were illustrated in this paper while taking them a stage further in selecting the best approximating model in the search for a realistic model of the dynamic process. Specifically, rather than attempt to estimate the empirical models only by assuming a reduced equation form for the initial outcome, which is sometimes a naïve form, we tested a variety of different flexible model equations, followed by a selection of the best model based on standard information criteria. In this paper, it is suggested that the pragmatic approach is preferred, in comparison with a variety of other approaches. It was shown that this approach dramatically improves the consistency of parameter estimation and precisely controls for the problem of negative estimates of variance components.

Acknowledgements

Authors would like to thank Marc Nerlove and Steve Bradley for their invaluable supports on earlier draft of this paper.

References

Aitkin, M., Alfò, M. (2003), "Longitudinal analysis of repeated binary data using autoregressive and random effect modelling", *Statistical Modelling,* Vol. 3 (4), pp. 291–303.

Anderson, T.W., Hsiao, C. (1981), "Estimation of dynamic models with error components", *Journal of the American Statistical Association,* Vol. 76, pp. 598–606.

Anderson, T.W., Hsiao, C. (1982), "Formulation and estimation of dynamic models using panel data", *Journal of Econometrics,* Vol. 18, pp. 47–82.

Baltagi, B.H. (1981), "Pooling: an experimental study of alternative testing and estimation procedures in a two-way error component model", *Journal of Econometrics,* Vol. 17, pp. 21–49.

Baltagi, B.H. (2001), *Econometric Analysis of Panel Data,* 2nd ed., John Wiley and Sons.

Bhargava, A., Sargan, J.D. (1983), "Estimating dynamic random effects models from panel data covering short time periods", *Econometrica,* Vol. 51, pp. 1635–1659.

Burnham, K.P., Anderson, D.R. (1998), *Model Selection and Inference: Practical Information-Theoretic Approach,* Springer, New York.

Breusch, T.S. (1987), "Maximum likelihood estimation of random effects models", *Journal of Econometrics,* Vol. 36, pp. 383–389.

Caselli, F., Esquivel, G., Lefort, F. (1996), "Reopening the convergence debate: a new look at the cross-country growth empirics", *Journal of Economic Growth,* Vol. 1, pp. 363–389.

Chamberlain, G. (1984), "Panel data", in: Griliches, Z., Intriligator, M., editors, *Handbook of Econometrics, Vol. 2,* North-Holland, Amsterdam, pp. 1247–1318.

Crouchley, R., Davies, R.B. (2001), "A comparison of gee and random effects models for distinguishing heterogeneity, nonstationarity and state dependence in a collection of short binary event series", *Statistical Modelling: An International Journal,* Vol. 1 (4), pp. 271–285.

Driscoll, P.J., William, P.F. (1996), "Presenting the Kuhn–Tucker conditions using a geometric approach", *The College Mathematics Journal*, Vol. 27 (2), pp. 101–108.

Hausman, J.A., Taylor, W.E. (1981), "Panel data and unobservable individual effects", *Econometrica*, Vol. 49, pp. 1377–1398.

Heckman, J.J. (1981), "The incidental parameter problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process", in: Manski, C.F., McFadden, D., editors, *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, MA, pp. 179–195.

Heckman, J.J. (1991), "Identifying the hand of past: distinguishing state dependence from heterogeneity", *The American Economic Review*, Vol. 81 (2), pp. 75–79.

Hsiao, C. (2002), *Analysis of Panel Data*, 2nd ed., University Press, Cambridge.

Hsiao, C., Pesaran, M.H., Tahmiscoioglu, A.K. (2002), "Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods", *Journal of Econometrics*, Vol. 109, pp. 107–150.

Islam, N. (1995), "Growth empirics: a panel data approach", *Quarterly Journal of Economics*, Vol. 110, pp. 1128–1170.

Kazemi, I., Davies, R.B. (2002), "The asymptotic bias of mles for dynamic panel data models", in: Stasinopoulos, M., Touloumi, G., editors, *Statistical Modelling in Society, Proceedings of the 17th IWSM, Chania, Greece*, pp. 391–395.

Kiviet, J.F. (1995), "On bias, inconsistency, and efficiency of various estimations in dynamic panel data models", *Journal of Econometrics*, Vol. 68, pp. 53–78.

Knight, M., Loayza, N., Villanueva, D. (1993), "Testing the neoclassical theory of economic growth", *IMF Staff Papers*, Vol. 40, pp. 513–541.

Lancaster, T. (2000), "The incidental parameter problem since 1948", *Journal of Econometrics*, Vol. 95, pp. 391–413.

Lee, K., Pesaran, M.H. (1998), "Growth empirics: a panel approach – a comment", *Quarterly Journal of Economics*, Vol. 113, pp. 319–323.

Lee, L.F. (1981), "Efficient estimation of dynamic error components models with panel data", in: Anderson, O.D., Perryman, M.R., editors, *Time-Series Analysis*, North-Holland, Amsterdam, pp. 267–285.

Maddala, G.S. (1987), "Recent developments in the econometrics of panel data analysis", *Transportation Research*, Vol. 21A, pp. 303–326.

Mankiw, N.G., Romer, D., Weil, D.N. (1992), "A Contribution to the empirics of economic growth", *Quarterly Journal of Economics*, Vol. 107, pp. 407–437.

McCulloch, C.E., Searle, S.R. (2000), *Generalized, Linear, and Mixed Models*, Wiley, New York.

Nerlove, M. (1999), "Likelihood inference for dynamic panel models", *Annales d'É conomie et de Statistique*, Vol. 5556, pp. 369–410.

Nerlove, M. (2000), "Growth rate convergence, fact or artifact?: an essay on panel data econometrics", in: Krishakumar, J., Ronchetti, E., editors, *Panel Data Econometrics: Future Directions (Papers in Honour of Professor Pietro Balestra)*, pp. 3–33.

Nerlove, M. (2002), *Essays in Panel Data Econometrics*, University Press, Cambridge.

Nerlove, M., Balestra, P. (1996), "Formulation and estimation of econometric models for panel data", Introductory essay in Mátyás and Sevestre, pp. 3–22.

Nickell, S. (1981), "Biases in dynamic models with fixed effects", *Econometrica*, Vol. 49, pp. 1417–1426.

Ridder, G., Wansbeek, T. (1990), "Dynamic models for panel data", in: van der Ploeg, R., editor, *Advanced Lectures in Quantitative Economics*, Academic Press, New York, pp. 557–582.

Sevestre, P., Trognon, A. (1996), “Dynamic linear models”, in: Mátyás, Sevestre, editors, *The Econometrics of Panel Data: A Handbook of the Theory with Applications 2nd Revised Edition*, Kluwer Academic Publishers, Dordrecht, pp. 120–144.

Summers, R., Heston, A. (1991), “The penn world table (mark 5). An expanded set of international comparisons, 1950–88”, *Quarterly Journal of Economics,* Vol. 106 (2), pp. 327–368.

This page intentionally left blank

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74005-6

CHAPTER 5

Time Invariant Variables and Panel Data Models: A Generalised Frisch–Waugh Theorem and its Implications

Jaya Krishnakumar

Department of Econometrics, University of Geneva, UNI-MAIL, 40, Bd. du Pont d'Arve, CH-1211, Geneva 4, Switzerland
E-mail address: jaya.krishnakumar@metri.unige.ch

Abstract

Mundlak ("On the pooling of time series and cross-section data", Econometrica, Vol. 46 (1978), pp. 69–85) showed that when individual effects are correlated with the explanatory variables in an error component (EC) model, the GLS estimator is given by the within. In this paper we bring out some additional interesting properties of the within estimator in Mundlak's model and go on to show that the within estimator remains valid in an extended EC model with time invariant variables and correlated specific effects. Adding an auxiliary regression to take account of possible correlation between the explanatory variables and the individual effects, we find that not only the elegant results obtained by Mundlak but also the above mentioned special features carry over to the extended case with interesting interpretations. We obtain these results using a generalised version of the Frisch–Waugh theorem, stated and proved in the paper. Finally, for both the EC models with and without time invariant variables we have shown that the estimates of the coefficients of the auxiliary variables can also be arrived at by following a two-step procedure.

Keywords: panel data, error components, correlated effects, *within* estimator

JEL classification: C23

5.1 Introduction

This paper is concerned with the issue of time invariant variables in panel data models. We try to look into an 'old' problem from a new angle or rather in an extended framework. It is well-known that when time invariant variables are present, the *within* transformation wipes them out and hence does not yield estimates for their coefficients. However they can be retrieved by regressing the means of the *within* residuals on these variables (see Hsiao, 1986, e.g.). Hausman and Taylor (1981) provide an efficient instrumental variable estimation of the model when the individual effects are correlated with some of the time invariant variables and some of the X's. Valid instruments are given by the other time invariant and time varying variables in the equation.

Suppose we consider the case in which the individual effects are correlated with all the explanatory variables. The earliest article dealing with this issue in panel data literature is that of Mundlak (1978) where the author looked at the error component model with individual effects and possible correlation of these individual effects with the explanatory variables (or rather their means). He showed that upon taking this correlation into account the resulting GLS estimator is the *within*. Thus the question of choice between the *within* and the random effects estimators was both "arbitrary and unnecessary" according to Mundlak.

Note that the question of correlation arises only in the random effects framework as the fixed effects are by definition non-stochastic and hence cannot be linked to the explanatory variables. We point this out because Mundlak's conclusion may often be interpreted wrongly that the fixed effects *model* is the correct specification. What Mundlak's study shows is that the estimator is the same (the *within*) whether the effects are considered fixed or random.

Now what happens to Mundlak's results when time invariant variables are present in the model? Do they still carry over? Or do they have to be modified? If so in what way? Are there any neat interpretations as in Mundlak's case? This paper is an attempt to answer these questions and go beyond them interpreting the results in a way that they keep the same elegance as in Mundlak's model.

The answers to the above questions follow smoothly if we go through a theorem extending the Frisch–Waugh result from the classical regression to the generalised regression. Thus we start in Section 5.2 by stating a generalised version of Frisch–Waugh theorem and giving its proof. In this section we also explain the important characteristic of this new theorem which makes it more than just a straightforward extension of the classical Frisch–Waugh theorem and point out in what way it is different from a similar theorem derived by Fiebig *et al.* (1996). The next section briefly recalls Mundlak's case and puts the notation in place. Section 5.4 brings out some interesting features of Mundlak's model which enable the known results. Section 5.5 presents the model with time invariant variables and discusses it from the point of view of correlated effects. Relationships between the different estimators are established and compared with the previous case. Finally we conclude with a summary of our main results.

5.2 The Generalised Frisch–Waugh theorem

THEOREM 5.1. *In the generalised regression model*:

$$y = X_1\beta_1 + X_2\beta_2 + u \tag{5.1}$$

with $E(u) = 0$ *and* $V(u) = V$*, positive definite, non-scalar, the GLS estimator of a subvector of the coefficients, say* β_2*, can be written as*

$$\hat{\beta}_{2,\text{gls}} = \left(R_2'V^{-1}R_2\right)^{-1}R_2'V^{-1}R_1, \tag{5.2}$$

where

$$R_1 = y - X_1\left(X_1'V^{-1}X_1\right)^{-1}X_1'V^{-1}y,$$

$$R_2 = X_2 - X_1\left(X_1'V^{-1}X_1\right)^{-1}X_1'V^{-1}X_2.$$

The proof of this theorem is given in Appendix A5.

Let us note an important property in the above formula for $\hat{\beta}_{2,\text{gls}}$ in that it represents a generalised regression of the residuals of GLS of y on X_1 on the GLS residuals of X_2 on X_1 with the *same initial* V as the variance covariance matrix in all the three regressions. An additional feature is that one can even replace R_1 by y in (5.2) and our result still holds (as in the classical case).

Fiebig *et al.* (1996) arrive at the GLS estimator $\hat{\beta}_2$ through a different route (applying M_1 to (5.1) and then (true) GLS on the transformed model). They also show that using a (mistaken) original V for their transformed model leads to a different estimator (which they call the pseudo GLS) and derive conditions under which pseudo GLS is equal to true GLS. Baltagi (2000) refers to Fiebig *et al.* (1996) while mentioning special cases examined by Baltagi and Krämer (1995, 1997) in which pseudo GLS equals true GLS.

Both our expression of $\hat{\beta}_2$ and Fiebig *et al.*'s (1996) true GLS expression yield the same answer but are obtained through different transformations. Expression (5.2) above has an interesting interpretation in terms of (GLS) residuals of auxiliary regressions as in the classical Frisch–Waugh case.

COROLLARY 5.1. *If in model (5.1) above we further have orthogonality between X_1 and X_2 in the metric V^{-1}, i.e. if*

$$X_1' V^{-1} X_2 = 0$$

then

$$\hat{\beta}_{1,\mathrm{gls}} = \left(X_1' V^{-1} X_1\right)^{-1} X_1' V^{-1} y,$$
$$\hat{\beta}_{2,\mathrm{gls}} = \left(X_2' V^{-1} X_2\right)^{-1} X_2' V^{-1} y.$$

5.3 The known case: Mundlak's model

Let us briefly recall Mundlak's result for a panel data model with only individual effects. The model is:

$$y = X\beta + (I_N \otimes \iota_T)u + w. \tag{5.3}$$

We have the usual assumptions $E(u) = 0$, $V(u) = \sigma_u^2 I_N$, $E(w) = 0$, $V(w) = \sigma_w^2 I_{NT}$ and independence between u and w. Thus denoting $\varepsilon = (I_N \otimes \iota_T)u + w$ we have $V(\varepsilon) \equiv \Sigma = \lambda_1 P + \lambda_2 Q$ with $\lambda_1 = \sigma_w^2 + T\sigma_u^2$, $\lambda_2 = \sigma_w^2$, $P = \frac{1}{T}(I_N \otimes \iota_T \iota_T')$ and $Q = I_{NT} - P$. Q is the well-known *within* transformation matrix.

When there is correlation between the individual effects u and the explanatory variables X, it is postulated using:

$$u = \overline{X}\gamma + v, \tag{5.4}$$

where $\overline{X} = \frac{1}{T}(I_N \otimes \iota_T')X$ and $v \sim (0, \sigma_v^2 I_N)$. Here one should leave out the previous assumption $E(u) = 0$. Substituting (5.4) into (5.3) we get

$$y = X\beta + (I_N \otimes \iota_T)\overline{X}\gamma + (I_N \otimes \iota_T)v + w. \tag{5.5}$$

Applying GLS to (5.5) Mundlak showed that

$$\begin{aligned} \hat{\beta}_{\mathrm{gls}} &= \hat{\beta}_w, \\ \hat{\gamma}_{\mathrm{gls}} &= \hat{\beta}_b - \hat{\beta}_w, \end{aligned} \tag{5.6}$$

where $\hat{\beta}_w$ and $\hat{\beta}_b$ are the *within* and the *between* estimators respectively.

Hence Mundlak concluded that the *within* estimator should be the preferred option in all circumstances.

5.4 Some interesting features

In this section we highlight some additional results for the above model which have interesting interpretations and lead us to the more general case of a model with time invariant variables.

Why within is GLS for β: Let us first look at the GLS estimation of the full model (5.5). Note that the additional term $(I_N \otimes \iota_T)\overline{X}$ can be written as PX.

Thus the augmented model becomes

$$y = X\beta + PX\gamma + \tilde{\varepsilon} \tag{5.7}$$

with $\tilde{\varepsilon} = (I_N \otimes \iota_T)v + w$ and $V(\tilde{\varepsilon}) \equiv \widetilde{\Sigma} = \tilde{\lambda}_1 P + \tilde{\lambda}_2 Q$ with $\tilde{\lambda}_1 = \sigma_w^2 + T\sigma_v^2$, $\tilde{\lambda}_2 = \sigma_w^2$.

Splitting X into its two orthogonal components QX and PX let us rewrite the above equation as

$$y = QX\beta + PX(\beta + \gamma) + \tilde{\varepsilon}. \tag{5.8}$$

Noticing that QX and PX are such that $X'Q\widetilde{\Sigma}^{-1}PX = 0$ we can apply Corollary 5.1 to obtain

$$\begin{aligned}\hat{\beta}_{\text{gls}} &= \left(X'Q\widetilde{\Sigma}^{-1}QX\right)^{-1} X'Q\widetilde{\Sigma}^{-1}y\\ &= (X'QX)^{-1}X'Qy = \hat{\beta}_w\end{aligned}$$

and

$$\begin{aligned}(\widehat{\beta + \gamma})_{\text{gls}} &= \left(X'P\widetilde{\Sigma}^{-1}PX\right)^{-1} X'P\widetilde{\Sigma}^{-1}y\\ &= (X'PX)^{-1}X'Py = \hat{\beta}_b.\end{aligned}$$

Thus we get back Mundlak's result (5.6):

$$\hat{\gamma}_{\text{gls}} = \hat{\beta}_b - \hat{\beta}_w.$$

This result can be further explained intuitively. Looking at model (5.7) we have X and PX as explanatory variables. Thus the coefficient of X, i.e. β measures the effect of X on y holding that of PX constant. Holding the effect of PX constant means that we are only actually measuring the effect of QX on y with β. Hence it is not surprising that we get $\hat{\beta}_w$ as the GLS estimator on the full model (5.7). However in the case of γ, it is the effect of PX holding X constant. Since X contains PX and QX as its components, we are only holding the QX component constant letting the PX component vary along with the PX which is explicitly in the equation whose combined effect is β and γ. Now the effect of PX on y is estimated by none other than the *between* estimator. So we have $(\widehat{\beta + \gamma})_{\text{gls}} = \hat{\beta}_b$, i.e. result (5.6) once again.

Within also equals an IV for β: As the X's are correlated with the error term $\varepsilon = (I_N \otimes \iota_T)u + w$, the GLS estimator will be biased but one could use instrumental variables. Various IV sets have been proposed in the literature (cf. Hausman and Taylor, 1981; Amemiya and McCurdy, 1986; Breusch *et al.*, 1989) and relative efficiency discussed at length. We will not go into that discussion here. Instead we point out that choosing the simple valid instrument QX also leads to the *within* estimator. Indeed, premultiplying Equation (5.3) by $X'Q$ we have

$$X'Qy = X'QX\beta + X'Q\varepsilon \tag{5.9}$$

and applying GLS we get the *within* estimator

$$\hat{\beta}_{IV} = (X'QX)^{-1}X'Qy = \hat{\beta}_w. \tag{5.10}$$

GLS for γ is equivalent to a two-step procedure: As far as γ is concerned, we observe that GLS on the full model is equivalent to the following two-step procedure:

Step 1: *Within* regression on model (5.3)
Step 2: Regression of *within* estimates of individual effects on $\overline{X}$ which gives $\hat{\gamma}$.

The individual effects estimates can be written as

$$\begin{aligned} u^* &= \frac{1}{T}(I_N \otimes \iota_T')\big[I_{NT} - X(X'QX)^{-1}X'Q\big]y \\ &= u + \frac{1}{T}(I_N \otimes \iota_T')\big[I_{NT} - X(X'QX)^{-1}X'Q\big]\varepsilon \end{aligned}$$

substituting (5.3) for y. Thus we have

$$u^* = \overline{X}\gamma + v + \frac{1}{T}(I_N \otimes \iota_T')\big[I_{NT} - X(X'QX)^{-1}X'Q\big]\varepsilon$$

or

$$u^* = \overline{X}\gamma + w^* \tag{5.11}$$

denoting $w^* = v + \frac{1}{T}(I_N \otimes \iota_T')[I_{NT} - X(X'QX)^{-1}X'Q]\varepsilon$.

It is interesting to verify that

$$V(w^*)\overline{X} = \overline{X}A$$

with A non-singular and hence we can apply OLS on (5.11). Thus we obtain

$$\begin{aligned} \hat{\gamma} &= (\overline{X}'\overline{X})^{-1}\overline{X}'u^* \\ &= (\overline{X}'\overline{X})^{-1}\overline{X}'(\bar{y} - \overline{X}\hat{\beta}_w) \\ &= \hat{\beta}_b - \hat{\beta}_w \end{aligned} \tag{5.12}$$

which is the same result as (5.6).

The above simple results not only show that we are able to arrive at the same estimator by various ways but also provide useful insight into the interesting connections working within the same model due to the special decomposition of the variance–covariance structure of EC models.

5.5 Extension to the case with time invariant variables

Now let us see what happens when time invariant variables come in. The new model is

$$y = X\beta + (I_N \otimes \iota_T)Z\delta + (I_N \otimes \iota_T)u + w = X\beta + CZ\delta + \varepsilon, \quad (5.13)$$

where Z is a $N \times p$ matrix of observations on p time-invariant variables relating to the N individuals and $C \equiv I_N \otimes \iota_T$.

5.5.1 Without correlated effects

Applying Theorem 5.1 on (5.13) and simplifying (see Appendix B5) one can obtain that $\hat{\beta}_{\text{gls}}$ is a weighted combination of the '*within*' and '*between*' (in fact an '*extended between*', see below) estimators, i.e.

$$\hat{\beta}_{\text{gls}} = W_1\hat{\beta}_{eb} + W_2\hat{\beta}_w, \quad (5.14)$$

where $\hat{\beta}_w$ is the same as before,

$$\hat{\beta}_{eb} = \left[X'\left(\frac{1}{T\lambda_1}CM_ZC'\right)X\right]^{-1}X'\left(\frac{1}{T\lambda_1}CM_ZC'\right)y \quad (5.15)$$

and W_1, W_2 are weight matrices defined in Appendix B5.

The estimator given in (5.15) is in fact the *between* estimator of β for an EC model with time invariant variables (as the *between* transformation changes the X's into their means but keeps the Z's as such; hence we have the transformation M_Z in between to eliminate the Z's). We call it the '*extended between*' estimator and abbreviate it as '*eb*'.

Turning to $\hat{\delta}_{\text{gls}}$, Theorem 5.1 implies

$$\hat{\delta}_{\text{gls}} = \left(F_2'\Sigma^{-1}F_2\right)^{-1}F_2'\Sigma^{-1}F_1, \quad (5.16)$$

where F_2 are residuals of CZ on X and F_1 are residuals of y on X. However for the former we should in fact be talking of residuals of Z on $\overline{X}$ as X is time varying and Z is time invariant. This means that in order to obtain $\hat{\delta}$ we should be regressing the individual *means* of residuals of y on X on those of Z on $\overline{X}$. Redefining F_1 and F_2 in this way and simplifying the expressions, we get

$$\begin{aligned}\hat{\delta}_{\text{gls}} &= (Z'M_{\overline{X}}Z)^{-1}Z'M_{\overline{X}}\frac{1}{T}(I_N \otimes \iota_T') \\ &\quad \times \left(I_{NT} - X\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}\right)y \\ &= (Z'M_{\overline{X}}Z)^{-1}Z'M_{\overline{X}}\bar{y}. \end{aligned} \quad (5.17)$$

5.5.2 *With correlated effects*

Now suppose that the individual effects are correlated with the X's and the Z's. The above estimators become inconsistent. Writing the auxiliary regression as

$$u = \overline{X}\gamma + Z\phi + v \tag{5.18}$$

and substituting u in (5.13) we get

$$\begin{aligned} y &= X\beta + CZ\delta + (I_N \otimes \iota_T)\overline{X}\gamma + (I_N \otimes \iota_T)Z\phi + (I_N \otimes \iota_T)v + w \\ &= X\beta + CZ(\delta + \phi) + PX\gamma + (I_N \otimes \iota_T)v + w. \end{aligned} \tag{5.19}$$

Within is still GLS for β: If we apply Theorem 5.1 to our model (5.19) above then we have the result that $\hat{\beta}_{\text{gls}}$ on (5.19) is the same as $\hat{\beta}_{\text{gls}}$ on the following model:

$$R_1 = R_2\beta + \varepsilon,$$

where

$$R_1 = y - \widetilde{Z}\left(\widetilde{Z}'\Sigma^{-1}\widetilde{Z}\right)^{-1}\widetilde{Z}'\Sigma^{-1}y$$

and

$$R_2 = X - \widetilde{Z}\left(\widetilde{Z}'\Sigma^{-1}\widetilde{Z}\right)^{-1}\widetilde{Z}'\Sigma^{-1}X$$

with

$$\widetilde{Z} = \left[\, (I_N \otimes \iota_T)Z \quad PX \,\right] = (I_N \otimes \iota_T)\,[\, Z \quad \overline{X}\,] = C\overline{Z}.$$

In other words,

$$\hat{\beta}_{\text{gls}} = \left(R_2'\Sigma^{-1}R_2\right)^{-1}R_2'\Sigma^{-1}R_1. \tag{5.20}$$

Once again making use of some special matrix results, one can show (see Appendix C5) that $\hat{\beta}_{\text{gls}} = \hat{\beta}_w$ for the augmented EC model with time invariant variables and correlated effects.

How can we intuitively explain this? Again it is straightforward if we write the model as

$$y = QX\beta + PX(\beta + \gamma) + CZ(\delta + \phi) + \varepsilon$$

and notice that QX is orthogonal to both PX and CZ in the metric Σ^{-1}. Corollary 5.1 above tells us that $\hat{\beta}_{\text{gls}}$ is given by

$$\hat{\beta}_{\text{gls}} = \left(X'Q\Sigma^{-1}QX\right)^{-1}X'Q\Sigma^{-1}Qy = (X'QX)^{-1}X'Qy = \hat{\beta}_w.$$

Within also equals an IV for β: Now it is easy to see that instrumenting X by QX in the new model (5.13) also leads to the *within* estimator for β coinciding with the GLS in the extended model. Of course transforming the model by the instrument matrix eliminates the time invariant variables just like the *within* transformation does. The coefficient estimates of the latter can always be retrieved in a second step by regressing the residual means on these same variables (see below).

GLS for γ is an 'extended' between–within: From the above intuitive reasoning we can also deduce that the parameters γ, δ and ϕ should be estimated together whereas we could leave out β as QX is orthogonal to both PX and Z in the metric Σ^{-1}.

Writing

$$\theta = \begin{bmatrix} (\delta+\phi) \\ (\beta+\gamma) \end{bmatrix}$$

we have by Theorem 5.1

$$\hat{\theta} = \begin{bmatrix} \widehat{(\delta+\phi)} \\ \widehat{(\beta+\gamma)} \end{bmatrix} = \big(\widetilde{Z}'\Sigma^{-1}\widetilde{Z}\big)^{-1}\widetilde{Z}'\Sigma^{-1}y.$$

Separate solutions for the two components of $\hat{\theta}$ can be obtained as yet another application of the same theorem:

$$\widehat{(\delta+\phi)} = (Z'M_{\overline{X}}Z)^{-1}Z'M_{\overline{X}}\bar{y},$$

$$\widehat{(\beta+\gamma)} = (\overline{X}'M_Z\overline{X})^{-1}\overline{X}'M_Z\bar{y},$$

where $\widehat{(\beta+\gamma)}$ can be recognised as the '*extended between*' estimator.[1] Once again the estimator of γ in the extended model is derived as the difference between the '*extended between*' and the *within* estimators:

$$\hat{\gamma}_{\text{gls}} = \widehat{(\beta+\gamma)} - \hat{\beta} = \hat{\beta}_{eb} - \hat{\beta}_{w}. \tag{5.21}$$

GLS for γ is again a two-step procedure: The above result on $\hat{\gamma}_{\text{gls}}$ leads to another interpretation similar to that of result (5.12) obtained in the model without time invariant variables. We have

$$\begin{aligned} \hat{\gamma}_{\text{gls}} &= (\overline{X}'M_Z\overline{X})^{-1}\overline{X}'M_Z\bar{y} - (X'QX)^{-1}X'Qy \\ &= (X'C'M_ZC'X)^{-1}X'CM_ZC'y - (X'QX)^{-1}X'Qy \\ &= (X'C'M_ZC'X)^{-1}X'CM_ZC'y \end{aligned}$$

[1] Here the '*between*' model is $\bar{y} = \overline{X}(\beta+\gamma) + Z(\delta+\phi) + \bar{\varepsilon}$.

$$
\begin{aligned}
&- (X'C'M_Z C'X)^{-1} X'CM_Z C'X(X'QX)^{-1}X'Qy \\
&= (X'C'M_Z C'X)^{-1} X'CM_Z C'\big(I_{NT} - X(X'QX)^{-1}X'Q\big)y \\
&= (X'C'M_Z C'X)^{-1} X'CM_Z C'\hat{u}^*
\end{aligned}
$$

which implies that $\hat{\gamma}_{\text{gls}}$ can be obtained by a two-step procedure as follows:

Step 1: *Within* regression of model (5.13).

Step 2: Regressing the *within* residual means on the residuals of the means of the X's on Z.

Now a few additional remarks. Note the formula for $(\widehat{\delta + \phi})$ is exactly the same as the one for $\hat{\delta}$ in the 'old' model (5.17) and this can be understood if we look into the effect captured by this coefficient. In model (5.13) δ is the effect of Z on y holding that of X constant, i.e. holding constant the effect of both the components QX and PX and the combined coefficient $(\delta + \phi)$ retains the same interpretation in the augmented model (5.19) too. However a major difference here is that one can only estimate the sum $(\delta + \phi)$ and cannot identify δ and ϕ separately. This is logical as both the coefficients are in a way trying to measure the same effect. Thus the inclusion of $Z\phi$ in the auxiliary regression (5.18) is redundant. The expression for $(\delta + \phi)$ can in fact be obtained by regressing $\hat{u}$ on $\overline{X}$ and Z. Thus, practically speaking δ and γ can be retrieved by regressing *within* residual means on $\overline{X}$ and Z.

Let us also mention that Hausman specification tests are carried out in the same manner whether time invariant variables are present or not and the absence of correlation can be tested using any one of the differences $\hat{\beta}_b - \hat{\beta}_w$, $\hat{\beta}_{\text{gls}} - \hat{\beta}_w$, $\hat{\beta}_{\text{gls}} - \hat{\beta}_b$ or $\hat{\beta}_{\text{gls}} - \hat{\beta}_{\text{ols}}$ as shown in Hausman and Taylor (1981).

If we assume non-zero correlation between explanatory variables and the *combined* disturbance term (the individual effects *and* the genuine disturbance terms), for instance in the context of a simultaneity problem, then the whole framework changes, *within* estimator is no longer consistent and only instrumental variables procedures such as the generalised 2SLS (G2SLS) or the error component 2SLS (EC2SLS) are valid (see, e.g., Krishnakumar, 1988; Baltagi, 1981).

5.6 Concluding remarks

In this paper we have shown that Mundlak's approach and the *within* estimator remain perfectly valid even in an extended EC model with time invariant variables. Adding an auxiliary regression to take account of possible correlation between the explanatory variables and the individual effects one finds that the elegant results obtained by Mundlak (1978) as well as some additional interesting ones can be derived in the extended case too. These results are established by the application of a generalised version of the Frisch–Waugh theorem also presented in the paper. Further, it is shown that for both the models with and without time invariant variables, the estimates of the coefficients of the auxiliary variables can also be obtained by a two-step estimation procedure.

Acknowledgements

The author would like to thank an anonymous referee, Pietro Balestra, Badi Baltagi, Geert Dhaene and the participants at the 2004 panel data conference for useful comments and suggestions.

Appendix A5

PROOF OF THEOREM 5.1. Let us transform the original model (5.1) by $V^{-1/2}$ to get

$$y^* = X_1^*\beta_1 + X_2^*\beta_2 + u^*,$$

where $y^* = V^{-1/2}y$, $X_1^* = V^{-1/2}X_1$, $X_2^* = V^{-1/2}X_2$ and $u^* = V^{-1/2}u$.

Now $V(u^*) = I_{NT}$ and hence we can apply the classical Frisch–Waugh theorem to obtain

$$\hat{\beta}_2 = \left(R_2^{*\prime} R_2^*\right)^{-1} R_2^{*\prime} R_1^*,$$

where

$$R_1^* = y^* - X_1^*\left(X_1^{*\prime} X_1^*\right)^{-1} X_1^{*\prime} y^*,$$

$$R_2^* = X_2^* - X_1^*\left(X_1^{*\prime} X_1^*\right)^{-1} X_1^{*\prime} X_2^*.$$

Substituting the starred variables in terms of the non-starred ones and rearranging we get

$$\begin{aligned}
\hat{\beta}_2 &= \left[X_2'\left(V^{-1} - V^{-1}X_1\left(X_1'V^{-1}X_1\right)^{-1}X_1'V^{-1}\right)X_2\right]^{-1} \\
&\quad \times X_2'\left(V^{-1} - V^{-1}X_1\left(X_1'V^{-1}X_1\right)^{-1}X_1'V^{-1}\right)y \\
&= \left[X_2'V^{-1}\left(I_{NT} - X_1\left(X_1'V^{-1}X_1\right)^{-1}X_1'V^{-1}\right)X_2\right]^{-1} \\
&\quad \times X_2'V^{-1}\left(I_{NT} - X_1\left(X_1'V^{-1}X_1\right)^{-1}X_1'V^{-1}\right)y \\
&= \left[X_2'\left(I_{NT} - V^{-1}X_1\left(X_1'V^{-1}X_1\right)^{-1}X_1'\right)V^{-1}
\end{aligned}$$

$$\begin{aligned}
&\times \left(I_{NT} - X_1\left(X_1'V^{-1}X_1\right)^{-1}X_1'V^{-1}\right)X_2\big]^{-1} \\
&\times X_2'\left(I_{NT} - V^{-1}X_1\left(X_1'V^{-1}X_1\right)^{-1}X_1'\right)V^{-1} \\
&\times \left(I_{NT} - X_1\left(X_1'V^{-1}X_1\right)^{-1}X_1'V^{-1}\right)y \\
= &\left(R_2'V^{-1}R_2\right)^{-1}R_2'V^{-1}R_1. \qquad \square
\end{aligned}$$

Appendix B5

Applying Theorem 5.1 on (5.13) yields:

$$\hat{\beta}_{\text{gls}} = \left(E_2'\Sigma^{-1}E_2\right)^{-1}E_2'\Sigma^{-1}E_1, \tag{B5.1}$$

where

$$\begin{aligned}
E_1 &= y - CZ\left(Z'C'\Sigma^{-1}CZ\right)^{-1}Z'C'\Sigma^{-1}y \\
&= \left(I_{NT} - \frac{1}{T}CZ(Z'Z)^{-1}Z'C'\right)y
\end{aligned}$$

and

$$\begin{aligned}
E_2 &= X - CZ\left(Z'C'\Sigma^{-1}CZ\right)^{-1}Z'C'\Sigma^{-1}X \\
&= \left(I_{NT} - \frac{1}{T}CZ(Z'Z)^{-1}Z'C'\right)X
\end{aligned}$$

using $C'\Sigma^{-1}C = \frac{1}{\lambda_1}TI_N$ and writing $\overline{X} = \frac{1}{T}C'X$.

Since $PC = C$, $QC = 0$, $CC' = TP$ and $C'C = TI_N$ one can see that

$$\begin{aligned}
\hat{\beta}_{\text{gls}} &= \left(E_2'\Sigma^{-1}E_2\right)^{-1}E_2'\Sigma^{-1}E_1 \\
&= \left[X'\left(\frac{\lambda_2}{T\lambda_1}CM_ZC' + Q\right)X\right]^{-1}X'\left(\frac{\lambda_2}{T\lambda_1}CM_ZC' + Q\right)y \\
&= W_1\hat{\beta}_{eb} + W_2\hat{\beta}_w,
\end{aligned}$$

where

$$\begin{aligned}
M_Z &= I_N - Z(Z'Z)^{-1}Z', \\
W_1 &= \left[X'\left(\frac{\lambda_2}{T\lambda_1}CM_ZC' + Q\right)X\right]^{-1}X'\left(\frac{\lambda_2}{T\lambda_1}CM_ZC'\right)X, \\
W_2 &= \left[X'\left(\frac{\lambda_2}{T\lambda_1}CM_ZC' + Q\right)X\right]^{-1}X'QX
\end{aligned}$$

and

$$\hat{\beta}_{eb} = \left[X'\left(\frac{1}{T\lambda_1}CM_ZC'\right)X\right]^{-1}X'\left(\frac{1}{T\lambda_1}CM_ZC'\right)y.$$

Appendix C5

We have from (5.20)

$$\hat{\beta}_{\text{gls}} = \left(R_2'\Sigma^{-1}R_2\right)^{-1}R_2'\Sigma^{-1}R_1. \tag{C5.1}$$

Let us examine R_1 and R_2. We can write them as $R_1 = \widetilde{M}y$ and $R_2 = \widetilde{M}X$ where $\widetilde{M} = I_N - \widetilde{Z}(\widetilde{Z}'\Sigma^{-1}\widetilde{Z})^{-1}\widetilde{Z}'\Sigma^{-1}$.

Noting once again that $PC = C$, $QC = 0$, $CC' = TP$, $C'C = TI_N$, $C'\Sigma^{-1}C = \frac{1}{\lambda_1}TI_N$, $\widetilde{Z}'\Sigma^{-1} = \frac{1}{\lambda_1}TZ'C'$ and $\widetilde{Z}'\Sigma^{-1}\widetilde{Z} = \frac{T}{\lambda_1}Z'Z$, one can show that $\widetilde{M} = I_{NT} - \frac{1}{T}C\overline{Z}(\overline{Z}'\overline{Z})^{-1}\overline{Z}'C' = I_{NT} - \frac{1}{T}CP_{\overline{Z}}C'$.

Further due to the partitioned nature of $\overline{Z}$ we also know that

$$P_{\overline{Z}} = P_{\overline{X}} + M_{\overline{X}}Z(Z'M_{\overline{X}}Z)^{-1}Z'M_{\overline{X}}.$$

Hence

$$\widetilde{M} = I_{NT} - \frac{1}{T}C\big(P_{\overline{X}} + M_{\overline{X}}Z(Z'M_{\overline{X}}Z)^{-1}Z'M_{\overline{X}}\big)C'$$

and

$$\widetilde{M}X = \left(I_{NT} - \frac{1}{T}C\overline{X}\right) = (I_{NT} - P)X = QX$$

as $P_{\overline{X}}C'X = TP_{\overline{X}}\overline{X} = T\overline{X} = C'X$ and $M_{\overline{X}}C'X = 0$. Therefore

$$R_2'\Sigma^{-1}R_2 = X'\widetilde{M}\Sigma^{-1}\widetilde{M}X = \frac{1}{\lambda_2}X'QX.$$

Similarly one can verify that

$$R_2'\Sigma^{-1}R_1 = X'\widetilde{M}\Sigma^{-1}\widetilde{M}y = \frac{1}{\lambda_2}X'Qy.$$

Thus

$$\hat{\beta}_{\text{gls}} = (X'QX)^{-1}X'Qy = \hat{\beta}_w.$$

References

Amemiya, T., McCurdy, T.E. (1986), "Instrumental variables estimation of an error components model", *Econometrica,* Vol. 54, pp. 869–881.

Baltagi, B.H. (1981), "Simultaneous equations with error components", *Journal of Econometrics,* Vol. 17, pp. 189–200.

Baltagi, B.H. (2000), "Further evidence on the efficiency of least squares in regression models", in: Krishnakumar, J., Ronchetti, E., editors, *Panel Data Econometrics: Future Directions,* North-Holland Elsevier, Amsterdam, pp. 279–291.

Baltagi, B.H., Krämer, W. (1995), “A mixed error component model”. Problem 95.1.4, *Econometric Theory*, Vol. 11, pp. 192–193.

Baltagi, B.H., Krämer, W. (1997), “A simple linear trend model with error components”. Problem 97.2.1, *Econometric Theory*, Vol. 13, p. 463.

Breusch, T.S., Mizon, G.E., Schmidt, P. (1989), “Efficient estimation using panel data”, *Econometrica*, Vol. 57, pp. 695–700.

Fiebig, D.G., Bartels, R., Krämer, W. (1996), “The Frisch–Waugh theorem and generalised least squares estimators”, *Econometric Reviews*, Vol. 15, pp. 431–444.

Hausman, J.A., Taylor, W.E. (1981), “Identification in linear simultaneous equations models with covariance restrictions: an instrumental variables interpretation”, *Econometrica*, Vol. 5 (51), pp. 1527–1549.

Hsiao, C. (1986), *Analysis of Panel Data*, Econometric Society Monographs, Cambridge University Press.

Krishnakumar, J. (1988), *Estimation of Simultaneous Equation Models with Error Components Structure*, Springer-Verlag, Berlin.

Mundlak, Y. (1978), “On the pooling of time series and cross-section data”, *Econometrica*, Vol. 46, pp. 69–85.

PART II

Empirical Applications

This page intentionally left blank

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74006-8

CHAPTER 6

An Intertemporal Model of Rational Criminal Choice

Robin C. Sickles[a] and Jenny Williams[b]

[a]Department of Economics, Rice University, 6100 South Main Street MS-22, Houston, TX 77005-1892, USA
E-mail address: rsickles@rice.edu
[b]Department of Economics, University of Melbourne, Melbourne, Vic 3010, Australia
E-mail address: jenny.williams@unimelb.edu.au

Abstract

We present a dynamic model of crime wherein agents anticipate future consequences of their actions. While investigating the role of human capital, our focus is on a form of capital that has received somewhat less attention in the literature, social capital, which accounts for the influence of social norms on the decision to participate in crime. The model is estimated with panel data from the 1958 Philadelphia Birth Cohort Study. Non-chosen states, which potentially depend on individual specific heterogeneity, are accounted for using simulation techniques. We find evidence of state dependence in the decision to participate in crime and the importance of initial levels of social capital stock in predicting adult criminal patterns.

Keywords: social capital, human capital, dynamic model, panel data, simulated method of moments

JEL classifications: C15, C33, C61, J22, Z13

6.1. Introduction

The basic premise of the economic model of crime is that criminals behave rationally in the sense that they act so as to maximize their economic welfare. This idea can be traced back to Bentham (1970 [1789]) and Beccaria (1963 [1764]), and has been more recently formalized by Becker (1968) and Ehrlich (1973). In this framework, a person breaks the law if the expected marginal benefit of allocating time and other resources to crime exceeds the marginal cost of doing so. To date, most empirical studies have focused on the labor market costs associated with criminal choice, investigating the effect of arrest history on current or future employment probabilities or wages.[1] However, recent theoretical and empirical research suggests that social interactions, working through peer influences, stigma, social norms, and information networks, also contribute to the cost and benefit calculations of many economic activities, including the decision to commit crime.[2] The role of social interactions is particularly relevant to the criminal participation decision if the stigma associated with arrest acts as a significant deterrent.

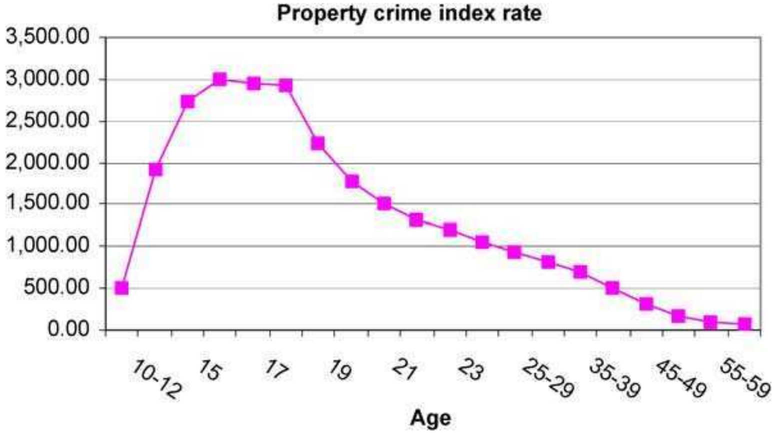
This research extends the traditional model of crime to explicitly account for the deterrent effect of social sanctions, or stigma, on the decision to participate in crime. We use social capital stock to measure an individual's past investment in the law-abiding social group, and assume that the cost of social sanctions faced depends upon the stock of social capital the individual has accumulated. In contrast to the literature on social capital that has followed in the tradition of Putnam, this study takes the level of social capital that a society possesses as given and, in the style of Coleman (1990), is concerned with the process by which individuals accumulate social capital stock and how this stock affects their behavior.[3] Our treatment of social capital as an individual characteristic is similar to Glaeser *et al*. (2002). However, this paper differentiates itself by its narrow focus on that part of social capital that embodies the norms associated with law-abiding behavior and the role of social capital in the enforcement of these norms. The intuition behind our approach is that attachment to (law-abiding) society through, for example, productive employment and marriage, creates a form of state dependence that reduces the likelihood of criminal involvement. In our formulation, state dependence arises because the stigma associated with arrest is more costly for individuals who

[1] See, for example, Imai and Krishna (2001), Lochner (1999), Grogger (1998, 1995), Waldfogel (1994), Freeman (1992), Lott (1990).

[2] See, for example, Akerlof (1997, 1998), Sampson and Laub (1992), Case and Katz (1991). The importance of the interaction between individuals and their community in forming tastes and determining criminal choices has been studied by Williams and Sickles (2002), Glaeser *et al*. (1996), Akerlof and Yellen (1994), and Sah (1991). The interaction between individuals decision to engage in crime and employers decision to stigmatize criminals is explored by Rasmusen (1996).

[3] The Putnam based social capital literature is interested in correlations between the level of social capital (proxied by measures of civic engagement, such as membership in organizations, and trust) that communities (measured at the state, regional and county level) have and outcomes such as good governance, economic growth or judicial efficiency (Putnam, 1993, 2000; Bowles and Gintis, 2002; Knack and Keefer, 1997; La Porta *et al*., 1997). As pointed out by Durlauf (2002), even within this genre, there is considerable ambiguity in what is meant by the term social capital.

Figure 6.1. Age specific Arrest rate[4]



have good jobs or families compared to those individuals without these attachments.

In addition to offering an explanation for differing criminal propensities across individuals, the model of social capital accumulation outlined in this paper provides a possible explanation for the age–arrest relationship. Figure 6.1 shows the age–arrest relationship for property arrests for the U.S. in 1999. The shape of this relationship, commonly called the age-crime profile, shows that the arrest rate increases with age up until the late teens, and then declines monotonically. This pattern has been found in studies based on different countries, cities and time periods. In our model, the relationship between age and arrest arises because it takes time to develop institutional relationships and hence accumulate social capital stock.[5] Therefore, crime becomes relatively more expensive and hence less likely for an individual as he ages.

Data from the 1958 Philadelphia Birth Cohort Study (Figlio *et al.*, 1991) are used to estimate our dynamic model of criminal choice. These data present a unique opportunity to study the dynamic decision to participate in crime. Typically, data used to study crime at the individual level are drawn from high-risk populations, such as prison releases, and consequently suffer from problems arising from selection bias. The data used

[4] The arrest rate is defined as the number of arrest per 100,000 in the population for each age. The data in Figure 6.1 are taken from Snyder (2000).

[5] Glaeser *et al.*'s (2002) model of investment in social capital predicts that social capital stock first rises and then declines with age, with the peak occurring at mid-life (around 50 years of age).

in this research are sampled from a universe of all individuals born in 1958 who lived in Philadelphia at least from their tenth until their eighteenth birthday. The information available in the Cohort Study includes direct measures of time spent working in the legal sector and both official and self-reported measures of involvement in crime. Secondary data sources are used to impute the time spent in crime based on the seriousness of offenses. Different criminal propensities arising from family background influences are accounted for by using these background variables in the construction of individual level initial values of social capital stock. The social capital stock accumulation process is then endogenously determined within the model, and the parameters governing this process are estimated within the system of Euler equations derived from the theoretical model.

An issue arising in estimation is that the ex-ante conditions for the optimality derived from the theoretical model depend on choices in each of two possible future states, arrest and escaping arrest. However, only one of these states will be realized and observed in the data. The presence of unobserved choices in the Euler equations pose an omitted regressor problem for estimation, and are potential source of unobserved heterogeneity. We address this issue using simulation techniques and estimate the parameters of our model by Method of Simulated Moments (McFadden, 1989; Pakes and Pollard, 1989; McFadden and Ruud, 1994).

The remainder of this paper is organized as follows. In the next section, we present a dynamic model of crime, which merges the intertemporal choice literature with Ehrlich's atemporal time allocation model of crime. Section 6.3 provides a description of the 1958 Philadelphia Birth Cohort Study and a discussion of the construction of our index of social capital stock. In Section 6.4 we discuss the method for estimating the structural parameters of the model and present the results from estimation. In Section 6.5, we offer some concluding remarks.

6.2. The model

In the spirit of Ehrlich (1973), we caste our model of criminal choice in a time allocation framework, where time represents the resources devoted to an activity. We extend this traditional static model to a dynamic setting by assuming that an individual's preferences and earnings depend upon his stock of social capital, which is a measure of his investment in the law-abiding group. In this model an individual's stock of social capital provides a flow of services associated with a good reputation and social acceptance within the law-abiding peer group, as well as social networks within this group. Reputation has utility value to the individual, while the networks can be used for occupational advancement and hence raise earnings in the legitimate sector.[6]

Consider the representative individual who must allocate his time between leisure λ_t, and the two income producing activities of legitimate work, L_t, and crime, C_t.[7] He must also choose his level of consumption X_t. At time t, utility is given by:

$$U(X_t, \lambda_t, S_t), \tag{6.1}$$

where S_t is the individual's stock of social capital. The utility function, $U(\cdot)$ is assumed to be twice differentiable, concave, and increasing in its arguments.

Denoting earnings within a period in terms of the composite good, X_t, the individual's intertemporal budget constraint is given by:

$$A_{t+1} = (1+r)\big(A_t + I_L(L_t, S_t) + I_C(C_t) - X_t\big), \tag{6.2}$$

where $I_L(L_t, S_t)$ is income from legitimate activity, $I_C(C_t)$ is income from illegitimate activity, and A_t represents the value of accumulated assets. We assume that per period income from legitimate work depends on the number of hours the individual spends working and the level of social capital he has accumulated. While a more general specification would allow both human and social capital stocks to influence earnings directly, including both in the structural model would increase the level of complexity for estimation because we could no longer obtain closed form solutions for the Euler equations.[8] In order to make the model tractable empirically, we focus on social capital in the theoretical model and control for standard measures of human capital, such as years of schooling and experience,

[6] Our model has several similarities with the model of social capital accumulation of Glaeser *et al*. (2002) in which the flow of services from social capital includes both market and non-market returns, where market returns may include higher wages or better employment prospects, and non-market returns may include improvements in the quality of the individual's relationships, improvements in health or direct happiness.

[7] In earlier work, both pure income and pure utility generating crimes were included in the model, where utility generating crime included rape and murder. However, the data did not contain sufficient information to identify the effect of utility generating crimes, so we have simplified the model by only considering income generating crimes.

[8] An approach to deal with this is to utilize asymptotic expansions to approximate the value function. In concert with the highly non-linear Euler equations system and the need to simulate unobserved states of apprehension/escape from apprehension, the additional computational burden of value function approximation is rather daunting. In this paper we concentrate on the social capital accumulation process in developing our theoretical structural dynamic model of crime while incorporating human capital indirectly into the empirical model.

in the empirical model.[9] The pecuniary rewards from income producing crime are assumed to depend only on the amount of resources devoted to this activity. This assumption is investigated in the empirical modeling of criminal earnings. Incomes from legitimate and illegitimate activities are assumed to be increasing in their respective arguments.

Investment in social capital is assumed to be proportionate to the level of resources spent in legitimate activity.[10] Resources in this model are represented by time. Social capital also depends on the state of the world. We assume that at the time the individual must choose how to allocate his time, he does not know if he will be arrested for participating in crime. This information is revealed at the end of the period. Thus, in the event of not being arrested (State 0) for crimes committed in time t, which occurs with probability $(1 - p)$, social capital at $t + 1$ is given by:

$$S^0_{t+1} = (1 - \delta)S_t + \gamma L_t, \tag{6.3}$$

where δ is the depreciation rate of social capital stock and γ transforms resources spent in legitimate activity into social capital. With probability, p, the individual will be arrested (State 1) at the beginning of $t + 1$ for crimes committed in time t and a social sanction imposed. This sanction is represented by a loss to the individual's social capital stock. We assume that this loss is an increasing function of the individuals stock of social capital so that, *ceteris paribus*, crime is more costly and therefore less likely for those with a greater stock in society. The loss is also assumed to depend positively on the total amount of time devoted to crime. Thus, in the event of apprehension, social capital at the beginning of $t + 1$ is given by:

$$S^1_{t+1} = (1 - \delta)S_t - \alpha C_t S_t, \tag{6.4}$$

where α represents the technology that transforms resources spent in crime into a social sanction.

A representative individual's dynamic programming problem is characterized by his value function at period t, $V(A_t, S_t)$, which is the solution to the Bellman equation:

[9] Formally, we are assuming that value function is a linear separable function of human capital.

[10] On the issue of investment in social capital, the approach taken in this paper differs from that taken by Glaeser *et al*. (2002) who assume that investing in social capital and work are mutually exclusive, and that the opportunity cost of investing in social capital is forgone earnings.

$$\begin{aligned}V(A_t, S_t) = \max_{X_t, L_t, C_t} & \; U(X_t, \lambda_t, S_t)\\ & + \beta\big\{pV\big(A_{t+1}, S^1_{t+1}\big) + (1-p)V\big(A_{t+1}, S^0_{t+1}\big)\big\}.\end{aligned}$$

Subject to (6.2), (6.3), (6.4) and a time constraint $T = \lambda_t + L_t + C_t$.[11] By substituting the time constraint in for λ_t, we eliminate it as a choice variable. Taking first-order conditions and making use of the Envelope Theorem, we obtain the following set of Euler equations:[12]

$$X_t\colon\; U_1(t) - \beta(1+r)\big\{pU^1_1(t+1) + (1-p)U^0_1(t+1)\big\} = 0,$$

$$\begin{aligned}L_t\colon\; & U_1(t)\frac{\partial I_L(L_t, S_t)}{\partial L_t} - U_2(t)\\ & + \beta\gamma(1-p)\left\{\left(\frac{(1-\delta)}{\gamma} - \left(\frac{1-\delta-\alpha C^0_{t+1}}{\alpha S^0_{t+1}}\right)\right)U^0_2(t+1)\right.\\ & + \left(\frac{\partial I_L(L^0_{t+1}, S^0_{t+1})}{\partial S_{t+1}} + \left(\frac{1-\delta-\alpha C^0_{t+1}}{\alpha S^0_{t+1}}\right)\frac{\partial I_C(C^0_{t+1})}{\partial C_{t+1}}\right.\\ & \left.\left. - \frac{(1-\delta)}{\gamma}\frac{\partial I_L(L^0_{t+1}, S^0_{t+1})}{\partial L_{t+1}}\right)U^0_1(t+1) + U^0_3(t+1)\right\} = 0,\end{aligned}$$

$$\begin{aligned}C_t\colon\; & U_1(t)\frac{\partial I_C(C_t)}{\partial C_t} - U_2(t)\\ & - \beta\alpha p S_t\left\{\left(\frac{(1-\delta)}{\gamma} - \left(\frac{1-\delta-\alpha C^1_{t+1}}{\alpha S^1_{t+1}}\right)\right)U^1_2(t+1)\right.\\ & + \left(\frac{\partial I_L(L^1_{t+1}, S^1_{t+1})}{\partial S_{t+1}} + \left(\frac{1-\delta-\alpha C^1_{t+1}}{\alpha S^1_{t+1}}\right)\frac{\partial I_C(C^1_{t+1})}{\partial C_{t+1}}\right.\\ & \left.\left. - \frac{(1-\delta)}{\gamma}\frac{\partial I_L(L^1_{t+1}, S^1_{t+1})}{\partial L_{t+1}}\right)U^1_1(t+1) + U^1_3(t+1)\right\} = 0,\end{aligned}$$

where $U^j_i(t+1)$ is the marginal utility of argument i ($i = 1, 2, 3$) in state j ($j = 0, 1$) at time $t+1$ and C^j_{t+1}, L^j_{t+1} represent choices in $t+1$ in state j.

The usual condition for optimality in consumption is given by the Euler equation for the aggregate consumption good, with the ratio of the marginal utility of current period consumption to the expected marginal utility

[11] An alternative formulation of the dynamic programming problem would include arrest status as a state variable. Using Theorem 4.2 of Stokey *et al.* (1989), Hartley (1996) shows that the solution to this problem will also solve the problem as formulated in the text.

[12] The derivation of the Euler equations can be obtained from the authors.

of next period's consumption equated to the gross real rate of interest. The Euler equation for time spent in the labor market equates net current period costs associated with time at work to the expected value of the increase in social capital in terms of next period decision variables. Similarly, the Euler equation for time spent in illegitimate income generating activities equates the net marginal benefit this period to the expected future cost. Once functional forms are specified for the utility and earnings functions, the system of three Euler equations and two earnings equations give a closed form solution for the optimal allocation of resources.

6.3. Data

We use individual level data drawn from the 1958 Philadelphia Birth Cohort Study to estimate the model developed in Section 6.2. Since these data have not had widespread use in economics literature, we begin with a description of the 1958 Philadelphia Birth Cohort Study and then discuss the sample used in the empirical part of the paper.

6.3.1. The 1958 Philadelphia Birth Cohort Study

The purpose of the 1958 Philadelphia Birth Cohort Study was to collect data on a birth cohort with a special focus on their delinquent and criminal activities. The cohort is composed of subjects who were born in 1958 and who resided in the city of Philadelphia at least from their tenth until their 18th birthday. The 27,160 members of this universe were identified using the Philadelphia school census, the U.S. Bureau of Census, and public and parochial school records. Once the members of this cohort were identified, data collection occurred in 2 phases.

The first phase of data collection involved assembling the complete official criminal history of the cohort. This was accomplished during the years 1979 and 1984 and provides coverage of the criminal careers, as recorded by the police, and juvenile and adult courts, for the entire 27,160 members of the cohort. The information for juveniles was obtained from the Philadelphia police, Juvenile Aid Division (JAD). Information about adult arrests was obtained from the Philadelphia Police Department, the Common and Municipal Courts, and the FBI, ensuring offenses both within and outside the boundaries of Philadelphia are included in the data set.

The second stage of the Study entailed a retrospective follow-up survey for a sample from the 27,160 members of the cohort. Figlio and his co-investigators employed a stratified sampling scheme to ensure that they captured the most relevant background and juvenile offense characteristics of the cohort and yield a sample size sufficient for analysis. The population was stratified five ways: by gender, race, socio-economic status, offense history (0, 1, 2–4, 5 or more offenses), and juvenile “status” offenses, which are offense categories only applicable to individuals less than 18 years of age. The follow-up survey took place during 1988, with 576 men and 201 women interviewed. Most respondents resided within the Philadelphia SMSA or within a 100-mile radius of the urban area. However, to insure that out-migration of cohort members from Philadelphia would not have any significant effect, sample members were traced and if possible contacted, throughout the United States. Figlio (1994) reports that comparisons among strata indicate no apparent biases due to non-response. Areas of inquiry covered by the survey include personal history of delinquency and criminal acts; gang membership; work and education histories; composition of current and childhood households; marital history; parental employment and educational histories; parental contact with the law; and personal, socioeconomic and demographic characteristics.

6.3.2. The sample

By combining the information from official arrest records with the retrospective survey data from the 1958 Philadelphia Birth Cohort Study, we have both self-reported information on criminal involvement and actual arrests, complete work histories, educational attainment, and a range of socio-economic and background characteristics for the sample captured in the retrospective survey. This paper focuses on males from the follow-up survey who were not full-time students so that leisure and work are the only alternatives to crime. We limit the sample to observations for which we can construct all key variables required to estimate the Euler equations derived from the theoretical model. Our final data set contains observations on 423 men over the ages of 19–24 corresponding to the period 1977 to 1982. A definition of variables and summary statistics are presented in Table 6.1.[13]

The choice variables from the structural model are (annual) hours spent in the labor market, (annual) hours spent in income producing crime, and (real) annual consumption. Income producing crimes are defined to be robbery, burglary, theft, forgery and counterfeiting, fraud, and buying, receiving or possessing stolen property. The annual number of hours worked in the legitimate labor market is constructed from the question, “How

[13] Since our data are from a stratified random sample, the statistics in Table 6.1 are calculated using weights to reflect the population from which the sample are drawn.

Table 6.1. Summary statistics

Definition	Mean	Standard deviation
Model variables		
Hours worked (L)	1498.04	934.61
Hours in income generating crime (C)	65.55	180.40
Leisure hours (λ)	4260.42	916.79
Real consumption per year (X)	119.23	84.65
Social capital index (S)	102.81	20.84
Real annual labor income (W_L)	100.69	91.83
Real annual crime income(W_C)	3.08	17.04
Determinants of social capital & earnings		
Binary equal to 1 if socio-economic status of family during childhood up is high	0.57	0.50
Binary equal to 1 if race is white	0.56	0.50
Binary equal to 1 if father present in childhood home	0.86	0.35
Binary equal to 1 if father not arrested during childhood	0.92	0.28
Binary equal to 1 if not a gang member during childhood	0.82	0.39
Number of siblings (divided by ten)	0.32	0.23
Proportion of best 3 friends not picked up by the police during high school	0.63	0.44
Number of police contacts as a juvenile	0.72	0.45
Proportion of contacts as a juvenile that result in an arrest	0.16	0.32
Binary equal to 1 if begin a marriage that year	0.05	0.21
Binary equal to 1 if end and then begin a job that year	0.10	0.30
Binary equal to 1 if arrested that year	0.05	0.22
Binary equal to 1 if arrested for a property offense that year	0.03	0.17
Binary equal to 1 if married	0.13	0.33
Binary equal to 1 if in a common law marriage	0.08	0.28
Number of children	1.00	1.13
Years of schooling	12.59	1.98
Years of labor market experience	1.52	1.68
Indicator for juvenile arrests	0.14	0.31

many hours per week did you usually work on this job?”, which was asked of each job recorded in the respondent’s work history. The Sellin–Wolfgang seriousness scoring scale is used to aggregate self-reported and official arrest information on crimes committed by the respondent each year (Sellin and Wolfgang, 1964). The seriousness score is then used to impute hours per year by matching the seriousness score to survey data recording hours spent in crime reported by Freeman (1992).[14]

[14] Details on the construction of these variables can be obtained from the authors. The sample used in estimation consists of 423 individuals and covers the years 1977–1982 (in-

In addition to the empirical counterparts to the variables in the structural model, Table 6.1 contains sample statistics for background characteristics that are used to construct the index of the initial level of social capital stock. These variables and the method used to construct this index are discussed later in this section.

6.3.3. Measuring social capital

6.3.3.1. Current social capital stock

We assume that gross investment in social capital in the sample period is generated by engaging in activities that develop institutional relationships such as attachment to the workforce and marriage. While providing detailed information on employment history, the 1958 Philadelphia Birth Cohort Study does not provide information on the level of involvement individuals have in their community. However, the Study does contain information about what Laub and Sampson (1993) and Sampson and Laub (1992) would consider turning points, such as marriage and beginning a new job. While much of the criminology literature has emphasized stability and continuity, Sampson and Laub argue that transitions are also important in understanding an individual's criminality, as these events may modify long-term patterns of behavior. For example, getting married forms social capital through a process of the reciprocal investment between husbands and wives. This investment creates an interdependent system of obligation and restraint and increases an individual's bonds to society. Also, young males tend to have high job turnover rates. If leaving a job and starting a new one in the same period is attributable to upward employment mobility, then a new job increases attachment to the legitimate sector when the employer's act of investing in the individual is reciprocated. Additionally, a better job increases an individual's system of networks. Each of these life events tends to increase an individual's ties to the legitimate community and thus increase his social capital.

In our empirical specification we follow the approach of Sampson and Laub, allowing getting married (GETMARRIED) and leaving and beginning a new job in the same period (CHANGEJOB) to build social capital stock. We account for stability of labor market attachment in our measure of social capital through annual hours spent in the legitimate labor market (L). Social capital also depends on the state of the world, which is

clusive) which corresponds to 2538 individual/year observations. Seriousness scores had to be generated for crimes for which there was no arrest. This amounts to 556 individual/years, which is about 22% of observations. The methodology used to accomplish this is available from the authors along with the aforementioned details on construction of variables.

learnt at the end of each period. In the event of not being arrested (State 0) for crimes committed in time t (C_t), social capital at $t+1$ is given by:

$$S^0_{t+1} = (1-\delta)S_t + \gamma_1 L_t + \gamma_2 \mathrm{GETMARRIED}_t + \gamma_3 \mathrm{CHANGEJOB}_t, \tag{6.5}$$

where δ is the depreciation rate of social capital and the γ's transform resources spent in legitimate activity into social capital.

Unlike legitimate income earning activities, criminal activity is not sanctioned by society. We model this by assuming that arrest results in a loss to the individual's social capital stock. As described in Section 6.2 the loss is assumed to depend positively on the resources devoted to crime and the level of social capital stock the individual has accumulated. Thus, in the event of apprehension, (State 1) social capital at $t+1$ is given by:

$$S^1_{t+1} = (1-\delta)S_t - \alpha C_t S_t, \tag{6.6}$$

where α represents the technology that transforms resources spent in crime into a social sanction. In order to estimate the weights $(\delta, \alpha, \gamma_1, \gamma_2, \gamma_3)$ in the capital accumulation process, we substitute Equations (6.5) and (6.6) in for S^0_{t+1} and S^1_{t+1} respectively in the Euler equations from Section 6.2. Once an initial level of social capital stock has been specified, these parameters can be estimated along with the other parameters of interest in the model.

6.3.3.2. Initial value of social capital stock

Since cohort members are eighteen at the beginning of our analysis, we assume that the initial period level of social capital stock possessed by an individual is inherited from his family. The choice of variables determining inherited social capital stock is based on empirical evidence from the literature, and the availability of these measures in our data. Becker (1991) notes that the fortunes of children are linked to their parents through endowments, such as family reputation and connections, knowledge, skills, and goals provided by the family environment. According to Coleman (1988), and the empirical literature on delinquency surveyed by Visher and Roth (1986), the institution of the family is central to the transmission of social norms to children and children's involvement in crime. Coleman notes that the creation of family bonds as a means of parents' instilling norms in their children depends not just upon the presence and willingness of the parents, but also on the relationship the children may have with competing norms and cultures, such as gang culture. Given our data, we account for each of these influences with the following variables: the socio-economic status of the individual's family during his childhood, race, whether the father was present in the childhood home, the number of siblings, whether the father was arrested during the individual's childhood, whether high school friends were in trouble with the police, gang membership during childhood, and the number of juvenile arrest relative to police contacts.

Obtaining a set of weights for aggregating variables such as presence of father, and gang affiliation during childhood raises the classic index number problem. Maasoumi (1986, 1993) shows that the (normalized) first principal component from the data on attributes can be used as weights to summarize these attributes into a composite index. In our application, we follow this approach.[15] We note that the use of principal components to initialize the stock of social capital is much like having a constant term in a human capital accumulation equation. We are interested in how changes in the stock of social capital impact changes in youth crime and these changes are determined within our model.

The variables with which we construct the initial stock of social capital are: father present in the childhood home, father not arrested during childhood, number of siblings, race, socioeconomic status, gang affiliation, proportion of best three friends from high school not picked up by the police, and the proportion of police contacts as a juvenile that result in arrests. The signs of the normalized weights associated with the first principal component indicate that coming from a white two-parent household with a high socioeconomic status, having a father with no arrests (during the individual's childhood), not being involved in a gang, and having friends who were not in trouble with the police contributes to the social capital stock an individual accumulates during childhood. The negative weight on the number of siblings indicates that the social capital stock a child inherits from his family is decreased by the presence of siblings. This is consistent with Coleman's (1988) finding that siblings dilute parental attention, which negatively effects the transmission of social capital from parents to child. Youths' involvement in criminal activity as measured by the ratio of juvenile arrests to police contacts also has a negative weight, indicating that juvenile arrests reduce the social capital stock accumulated during childhood. Inherited social capital is constructed as the weighted sum of these variables.

[15] These weights are sample specific. As an alternative, Maasoumi (1986, 1993) suggests that the weights given to the attributes may be the researcher's subjective weights. Factor analysis is an alternative means to obtain weights. However, Kim and Mueller (1978) note that principal components has an advantage over factor analysis if the objective is a simple summary of information contained in the raw data, since the method of principal components does not require the strong assumptions underlying factor analysis.

The index of inherited social capital stock should provide a measure of the degree to which an individual is “at risk” of criminal involvement and arrest in the sample period. Specifically, we would expect that individuals with a smaller stock spend more time in crime and are more likely to be arrested than individuals who inherited a larger stock. We investigate whether this is the case by dividing the sample into quartiles based on the initial level of social capital stock and comparing the first and fourth quartiles in terms of two measures of criminal involvement: arrests and time in crime. Individuals from the first quartile of inherited social capital stock account for a much larger proportion of annual arrests for the sample than men from the fourth quartile, and this difference becomes more pronounced over time. Moreover, those from the first quartile of social capital stock inherited from the family do spend a much larger amount of time in crime relative to those from the fourth quartile. A t-test for the equality of means (allowing for unequal variances) between the first and fourth quartiles indicates a significant difference for each year. This confirms that the initial level of social capital stock is a good predicator of propensity for criminal involvement in adulthood.

6.4. Empirical model

The Euler equations derived from the structural model of crime in Section 6.2 depend on state contingent choices in each of two possible future states, apprehension and escaping apprehension. However, only one of these future states will be realized and observed in the data. The unobserved choices cause an omitted regressor problem in estimation and are a potential source of unobserved heterogeneity. While it is possible to estimate the three Euler equations and two income equations simultaneously, the absence of unobserved choices in the earnings equations makes a sequential estimation process computationally convenient. However, because the parameters governing social capital accumulation are estimated from the Euler equations, and are then used to construct the social capital stock that enter into the earnings equations, the estimation algorithm iterates between earnings and Euler equation estimation.

In terms of describing our estimation strategy, we begin with describing estimation of the parameters in the earnings equations, which draws on standard techniques in the labor econometrics literature. Section 6.4.2 describes the method for estimating the parameters of the utility function and social capital accumulation function from the Euler equations, which is based on the Method of Simulated Moments (McFadden and Ruud, 1994; McFadden, 1989; Pakes and Pollard, 1989).

6.4.1. The earnings equations

6.4.1.1. Estimation methodology for the earnings equations

The model presented in Section 6.2 focuses on the role of social capital in decisions regarding participation in crime and work. This leads to a specification for criminal earnings that depends on resources the individual allocates to that activity, and legitimate labor market earnings that depends on both hours spent working and social capital stock. However, in addition to the large empirical literature on human capital, empirical research by Freeman (1996) suggests that the return to legitimate opportunities relative to the returns to crime also depends on human capital. Further, he finds that human capital affects relative income through raising the return to work. To reflect this in our empirical model, we adopt a more general specification that includes human capital as a determinant of legitimate earnings. We also explore whether criminal human capital (and legitimate human capital) raises the returns to time in crime.

Income in each sector is defined as the product of the number of hours spent in that sector and that sector's hourly wage:

$$I_L = w_L(H_t, S_t, Z_t) \cdot L_t,$$

$$I_C = w_C(K_t, Z_t) \cdot C_t,$$

where w_L and w_C are the hourly wage in the legitimate labor market and criminal labor markets respectively. L_t and C_t denote hours per year in legitimate and criminal income generating activities respectively, S_t is the social capital stock accumulated by the individual at the beginning of period t, H_t is legitimate human capital, represented by years of schooling and labor market experience, K_t is criminal human capital, and Z_t represents a vector of socioeconomic and demographic characteristics including marital status, number of children and race. We measure criminal human capital stock using the number of juvenile arrests (as a proxy for experience) and a variable indicating whether the respondent's father was arrested in the respondent's youth and a variable measuring the respondent's number of siblings (as a proxy for criminal networks).

The wage equations are intended to provide us with information about the determinants of wages for the entire sample of men. However, the decision to participate in each sector is endogenous, and only a sub-sample of the population is engaged in either or both of the income producing activities. If the decision to work in legitimate or illegitimate activities depends on unobservable characteristics that also influence wages, then the problem of sample selection exists. Since we are estimating the earnings equations separately from the Euler equations, we make use of standard econometric techniques to account for the possibility of sample selection bias (Heckman, 1974, 1979).

6.4.1.2. Earnings equation results

The estimates for the sample selection corrected wage equations for criminal and legitimate activities are presented in Table 6.2.[16] Hourly wages in the legitimate labor market are constructed by linear interpolation between the reported pay the individual received when they started and left each job in their employment history. If earnings were reported as weekly (yearly), the hourly wage is calculated as the weekly (yearly) wage divided by the usual hours worked per week (usual hours worked per week multiplies by 50 weeks). Annual criminal income is defined as the total value of stolen goods from arrests and self-reported offenses. The hourly wage for property crime is then calculated as the annual income divided by the number of hours spent in crime that year.[17]

The parameter estimates for the legitimate labor market wages equation are consistent with the standard predictions of human capital theory. Legitimate wages are increasing in years of schooling, and are a concave function of labor market experience. In addition to the human capital theory of earnings, we find evidence that institutional knowledge and networks, as captured by our measure of social capital stock, has a positive and significant impact on earnings. These results suggest that both human capital and social capital are significant determinants of wages.

In contrast to labor market wages, we are unable to explain criminal wages with criminal human capital variables, nor are we able to explain criminal wages with the legitimate human capital measures. The joint hypothesis that criminal (legitimate) human capital and the socioeconomic and demographic variables are insignificant in explaining criminal wages cannot be rejected at conventional levels of significance, with a p-value for the Wald test statistic of 0.59 (0.57). This may reflect problems with measuring criminal income, hours, or criminal human capital. Alternatively, the finding may reflect that criminal earnings are not related to either legitimate or criminal human capital. We note that while not significant in

[16] We used a pooled regression to estimate the hourly wage and participation equation. We were unable to utilize a fixed effects estimator because of time invariant regressors. The time invariant regressors identify the model and their inclusion is therefore necessary. A random effects estimator is an alternative that could accommodate the time invariant regressors. Both the random effects estimator and the estimator used provide consistent point estimates under the assumption that the effects are uncorrelated with included regressors. The key objective of estimating the wage equations is to obtain consistent estimates of the equation parameters in order to estimate the Euler equations and the method used achieves this end. The results are used to calibrate our simulated GMM model presented in Section 6.4.2 below.

[17] A full description of the construction of this variable can be obtained from the authors.

Table 6.2. Selection corrected equations for hourly wages in work and crime

	Work		Crime		Crime	
Log hourly wage	Parameter	*t*-value	Parameter	*t*-value	Parameter	*t*-value
Years of schooling	0.026	3.008	−0.086	−1.228	−0.079	−1.069
Experience	0.069	2.574			0.240	1.288
Experience squared	−0.009	−2.267			−0.051	−1.465
Father arrested during respondent’s childhood			−0.157	−0.524		
Number of juvenile arrests			−0.153	−0.419		
Number of siblings			−0.072	−1.577		
Social capital	0.001	2.138	0.001	0.100	0.002	0.218
Race is white	0.057	2.185	0.058	0.186	0.104	0.322
Indicator for married	0.025	0.865	0.288	0.887	0.219	0.668
Indicator for in a common law marriage	0.088	2.477	−0.281	−0.825	−0.268	−0.734
Year	−0.045	−4.446	−0.042	−0.515	−0.042	−0.485
Constant	0.338	0.448	1.308	0.204	0.626	0.093
p-value of Wald test for joint significance of regressor		0.000		0.592		0.565

(*continued on next page*)

Table 6.2. (Continued)

	Work		Crime		Crime	
Participation	Parameter	*t*-value	Parameter	*t*-value	Parameter	*t*-value
Years of schooling	0.153	4.530	−0.034	−1.265	−0.034	−1.271
Experience	1.020	12.614	−0.127	−2.007	−0.128	−2.073
Experience squared	−0.116	−7.378	0.005	0.417	0.005	0.458
Social capital	0.008	2.631	−0.017	−7.319	−0.017	−7.299
Race is white	0.257	2.605	0.442	5.225	0.442	5.223
Indicator for married	0.543	3.148	−0.002	−0.021	−0.004	−0.034
Indicator for in a common law marriage	0.175	1.303	0.545	5.065	0.546	5.063
Number of children	0.032	0.997	−0.040	−1.417	−0.041	−1.420
Moved out of parents home	−0.027	−0.161			0.031	0.223
Father was arrested	−0.375	−2.944	0.248	2.235	0.247	2.218
Number of juvenile arrests	−0.270	−1.975	0.373	3.655	0.373	3.631
Number of siblings	−0.035	−1.676	−0.009	−0.544	−0.009	−0.553
Year	−0.150	−4.200	−0.017	−0.566	−0.016	−0.535
Constant	9.418	3.335	2.533	1.079	2.473	1.045
p-value of Lagrange Multiplier test for independent equations		0.963		0.931		0.941

determining wages, two out of three measures of criminal human capital (number of juvenile arrests and father was arrested in respondent's youth) are significant in explaining participation in crime, as is martial status, and social capital, with participation less likely at higher levels of social capital stock. While we cannot rule out measurement issues as the reason for being unable to explain criminal wages, we note that Freeman (1996) finds that human capital affects relative income through raising returns to legitimate work rather than through criminal income.[18] Also Gottfredson and Hirschi (1990) concluded that for the vast majority of income generating crimes such as theft and burglary, there is no evidence of criminal human capital accumulation. From the combined evidence, it may be reasonable to infer that criminal returns are not a function of criminal human capital.

As we are unable to explain criminal wages with human capital, criminal capital, or socioeconomic and demographic variables, we adopt the assumption used in the theoretical model that criminal income depends on time spent in crime only. Accordingly, we estimate a criminal income function as follows:

$$W_C(C_t) = \mu_0 + \mu_1 C_t + \mu_2 C_t^2 + \varepsilon_{C_t}.$$

Since time in crime is a choice variable potentially correlated with the error term in the earnings equation, and is truncated below by zero, we correct for the potential for sample selection bias by adopting the methodology suggested in Vella (1998). This approach is similar to the parametric two-step approach of Heckman (1974, 1979). In the first step, we assume normality of the error term in the latent variable reduced form equation for hours worked, leading to a Tobit specification. However, distributional assumptions about the error term in the earnings equation are relaxed in the second step. This leads us to approximate the selection term in the earnings equation by $\sum_{k=1}^{K} \alpha_k \hat{\nu}_k^k$ where the $\hat{\nu}_k$ are the generalized residuals from the first-step Tobit estimation and K is the number of terms in the approximating series. By including this polynomial in the earnings equation, we take account of the selection term. Therefore, exploiting the variation in hours worked (in illegitimate income producing activities) for the subsample that participates provides consistent OLS estimates of parameters in the criminal earnings equation. Provided K is treated as known, these estimates are $\sqrt{n}$ consistent, and the second step covariance matrix can be computed.

[18] Specifically, he regressed the share of income from illegal sources on human capital measures and found that the coefficients on all human capital variables were negative and significant.

Table 6.3. Selection corrected criminal annual earnings equation

	Parameter	t-value
Criminal earnings		
Hours in crime	9.01×10^{-3}	0.560
Hours in crime squared	5.13×10^{-6}	3.190
Constant	7.718	4.050
Hours in crime		
Years of schooling	−14.546	−1.303
Experience	−48.416	−1.878
Experience squared	2.450	0.506
Social capital	−7.648	−7.896
Race is white	128.467	3.685
Indicator for married	31.115	0.677
Indicator for in a common law marriage	174.480	4.076
Number of children	−24.820	−2.109
Moved out of parents home	1.273	0.023
Father was arrested	169.382	3.852
Number of juvenile arrests	122.220	2.822
Number of siblings	−0.566	−0.082
Year	−13.947	−1.115
Constant	1648.632	1.679

The results from estimating the sample selection corrected criminal earnings function are presented in Table 6.3. The results are from an OLS regression whose standard errors are consistent under the null hypothesis that the residual terms are jointly insignificant which we find is the case.[19] We examined different treatments of pooling in the earnings equation but were unable to identify the coefficients with a within type estimator. Results with an error components specification were quite similar to the OLS results. These estimates are used to calibrate the Euler equations in the simulated GMM estimation and are not the focus per se of our empirical model. Results are in line with findings in other studies of earnings. Annual income from crime is an increasing function of time spent in that activity. Increasing returns to time in crime may be evidence of some fixed cost, or accumulation of crime specific networks and knowledge.

Given there appear to be increasing returns to time in crime we would expect individuals who participate in crime to specialize. However, eighty percent of men in our sample who engage in crime also work in the legitimate sector. Among criminals who do work, an average of one and

[19] The p-value of F test for joint insignificance of correction terms is 0.740. We set $K = 3$.

one-half hours per week is spent in crime compared to almost 36 hours per week working at a legitimate job. This implies there are costs associated with crime, or benefits associated with not engaging in crime, which are not captured by the earnings equations. According to our model, these benefits are the utility value of social capital, such as social acceptance and reputation, representing state dependence in non-deviant behavior in the preference structure. We investigate this hypothesis in the next section by estimating the Euler equations associated with the optimal allocation of time to criminal and legitimate activities, and consumption.

6.4.2. The Euler equations

6.4.2.1. Estimation methodology for the Euler equations

Let S_{it} denote the value of the state variable, social capital stock, for the ith individual in period t, x_{it} denote the vector of choice variables entering the ith individual's Euler equations in period t, and let x_{it+1} be those variables dated $t+1$. Our sample is a panel of $T = 5$ periods of observations on a random sample of $N = 423$ individuals. We assume that the earnings in the legal and criminal sectors are parameterized as above and that utility has the following transcendental logarithmic form:

$$\begin{aligned} U(X_{it}, \lambda_{it}, S_{it}) = \alpha_1 \ln X_{it} + \alpha_2 \ln \lambda_{it} + \alpha_3 \ln S_{it} + \frac{1}{2}\big\{\beta_{11}(\ln X_{it})^2 \\ + \beta_{22}(\ln \lambda_{it})^2 + \beta_{33}(\ln S_{it})^2\big\} + \beta_{12} \ln X_{it} \ln \lambda_{it} \\ + \beta_{13} \ln X_{it} \ln S_{it} + \beta_{23} \ln \lambda_{it} \ln S_{it}. \end{aligned}$$

Each of these Euler equations can be written in the form of $f_j(x_{it}, S_{it}, \theta_0) - g_j(x_{it+1}, S_{it+1}, \theta_0)$, $j = 1, 2, 3$, where $f(\cdot)$ is the observed response function which depends on current period variables, and $g(\cdot)$ is the expected response function, which depends on next period's variables, and θ_0 is the $p \times 1$ vector of parameters to be estimated.[20] A stochastic framework is introduced by assuming that variables determined outside the model, whose future values are unknown and random, cause agents to make errors in choosing their utility maximizing bundles. The errors u_{it} are idiosyncratic so that at any time, the expectation of this disturbance term over individuals is zero. The ith individual's system of equations is represented as:

$$f(x_{it}, S_{it}, \theta_0) - g(x_{it+1}, S_{it+1}, \theta_0) = u_{it}.$$

[20] We assume a real rate of interest of 3%, and a time rate of preference of 0.95. The representative individual's per period optimal choice of time allocations (L_t, C_t) and consumption (X_t) are parameterized by $\theta_0 = (\alpha_1, \alpha_2, \alpha_3, \beta_{11}, \beta_{22}, \beta_{33}, \beta_{12}, \beta_{13}, \beta_{23}, \alpha, \delta, \gamma_1, \gamma_2, \gamma_3)$.

Conditional moment restrictions take the form, $E[u_{it}|z_{it}] = 0$, where z_{it} are observed data.

In practice, implementing GMM as an estimator for the parameters in our system of Euler equations is hampered by the fact that, while an agent's decision is based on ex-ante expectations of the future, ex-post only one state is realized for each individual and subsequently observed by the econometrician. Since the (unobserved) choice in the state not realized enters the Euler equations through $g(x_{it+1}, S_{it+1}, \theta_0)$, we are faced with an omitted regressor problem in the expected response function. We resolve this problem by replacing $M(\cdot)$ with a simulator, $\mu(\cdot)$. McFadden (1989) proposes this modification of the conventional Method of Moments estimator as the basis for the Method of Simulated Moments.[21]

To illustrate our use of MSM, recall that individual i's current choice x_{it} depends on the value of the state variable, social capital stock, S_{it}. Our problem is that x_{it+1} is not observed for individual i in the state not realized in period $t+1$, so sample averages of $M(\cdot)$ cannot be formed. However, if the density, $\Pi(x, S)$, is stationary then we can replace the unobserved x_{it+1} with Monte Carlo draws from the conditional distribution, $\Pi(x|S_{t+1})$. Recall that S_{t+1} depends on last period's choices, and whether or not the individual is apprehended in period $t+1$, so we are able to construct future social capital stock in period $t+1$ in the unobserved state for a given set of parameters governing social capital accumulation. Since this distribution is unknown, we draw from the empirical conditional distribution, which is estimated by kernel-based methods. Having replaced the unobserved data with the Monte Carlo draws, we then form a simulator of our moment conditions as follows:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{S}\sum_{s=1}^{S}\big(f(x_{it}, S_{it}, \theta_0) - g\big(x_{it+1}^{s}, S_{it+1}, \theta_0\big)\big) \otimes z_{it}\right]\\ = \mu(x_i, S_i, z_i, \theta_0),\end{aligned}$$

where

$$\lim_{N\to\infty} E_N\left[\frac{1}{N}\sum_{i=1}^{N}\big[\mu(x_i, S_i, z_i, \theta_0)\big]\right] = E_N\big[M(x_i, S_i, z_i, \theta_0)\big].$$

[21] Sufficient conditions for the MSM estimator to be consistent and asymptotically normal involve the same regularity assumptions and conditions on instruments as classical GMM, in addition to the two following assumptions that concern the simulator, $\mu(\cdot)$: (i) the simulation bias, conditional on W_0 and x_{it}, is zero, and (ii) the simulation residual process is uniformly stochastically bounded and equicontinuous in θ.

Table 6.4. Estimates of structural parameters from Euler equation estimation

	Parameter	t-value
Translog utility function parameters		
$\ln X_t$	0.2258	2.09
$\ln \lambda_t$	0.2060	0.47
$(\ln X_t)^2$	0.0028	2.61
$(\ln \lambda_t)^2$	0.1069	2.09
$(\ln S_t)^2$	0.1908	2.85
$\ln X_t \ln \lambda_t$	−0.0179	−1.46
$\ln X_t \ln S_t$	−0.0160	−6.31
$\ln S_t \ln \lambda_t$	−0.2141	−6.61
Social capital accumulation parameters		
δ	0.0299	2.23
$\gamma 1$	0.0003	0.64
$\gamma 2$	4.0800	1.37
$\gamma 3$	15.1400	1.76
α	0.0002	0.67

Note that although we motivate the estimation methodology as a way of dealing with uncertainty about future states, the use of simulation techniques conditioned on individual characteristics may also be viewed as a partial control for unobserved individual heterogeneity in those states.

6.4.3. Euler equation results

The system of Euler equations derived in Section 6.2 is estimated using MSM on 423 individuals over the period 1977 to 1981. The coefficient on the logarithm of social capital (α_3) is normalized at unity, leaving eight coefficients from the translog utility function and five parameters from the social capital accumulation process to be estimated. With three equations and eleven instruments, the number of overidentifying restrictions is twenty. The Hansen test statistic for overidentifying restrictions is 6.65, compared to a $\chi^2_{0.95,20} = 10.85$ so the null hypothesis that the system is over-identified is not rejected. The MSM estimates of the preference parameters are presented in the top half of Table 6.4, and the parameters governing the accumulation of social capital stock in the bottom half of this table. It is noteworthy that all three terms in the translog utility function involving social capital are significantly different from zero, supporting the hypothesis that preferences exhibit state dependence.

Examining the estimates of the translog preference parameters in Table 6.4, we find the coefficients on the interaction terms between consumption and leisure ($\ln X_t \ln \lambda_t$), consumption and social capital ($\ln X_t \ln S_t$), and leisure and social capital ($\ln \lambda_t \ln S_t$) are all negative. Our estimates imply that consumption and leisure are complements in utility. This is consistent with the work of Hotz *et al*. (1988), Sickles and Taubman (1997), and Sickles and Yazbeck (1998).[22] The relationships between consumption and social capital, and leisure and social capital, are also complementary. Moreover, these interaction terms are statistically significant.

Turning to the parameters governing social capital accumulation, we estimate a statistically significant depreciation rate on social capital stock (δ) of 3%. The sign on the point estimates of time in the labor market (γ_1), getting married (γ_2), and changing jobs (γ_3) are all positive, indicating that they each contribute to social capital stock accumulation, although only γ_2 and γ_3 are statistically significant at the 10% level of significance using a one-sided test. While not statistically significant, the coefficient on the social penalty for arrest (α) implies a loss of 1% of social capital stock evaluated at the sample average of time in crime. Evaluated at the mean annual hours spent in crime amongst the criminally active, the social sanction is about 5% of social capital stock.

Returning to the preference parameters, we note that the estimated marginal utilities of consumption, leisure, and social capital are positive for all time periods.[23] The value of an incremental increase in the consumption good drops from ages 19 to 20, and rises from the age of 20 for our sample of young men. The marginal utility of leisure declines steeply between the ages of 19 and 20, continues to decline between the ages of 20 and 21, and then increases over the ages of 21–23. Based on these estimates, the average marginal rate of substitution of consumption for leisure is 0.056, implying an hourly wage of $4.18 over the sample period.[24] The marginal rate of substitution of consumption for leisure is about an order of magnitude smaller than the value of 0.8667 obtained by Sickles and Yazbeck (1998), who use data from the Retirement History Survey. This may be evidence that older individuals place a higher value on leisure time.

The marginal utility of social capital also increases over time for our sample of young men. In addition to growing state dependence, this result indicates that agents are indeed forward looking in their decision-making. Over the sample period, average leisure time decreases as individuals spend a greater amount of time in employment. Current labor market

[22] Other studies, however, find evidence that these goods are substitutes (Altonji, 1986; Ghez and Becker, 1975; Thurow, 1969).

[23] These are obtained by evaluating at sample averaged (across individuals) data.

[24] This number is calculated by multiplying the marginal rate of substitution by the CPI, where the CPI is averaged over 1977 to 1981.

activity is expected to increase future welfare through social capital accumulation, and this in turn raises the marginal utility of social capital in the current period. Thus, the marginal utility of past investment in social capital is increasing in current investment. Alternately, the marginal utility of current investment in social capital is increasing in past investment. This is a necessary condition for adjacent complementarity.[25] Since past labor market participation raises social capital stock, which raises future labor supply, we also find reinforcement in decision-making.

To gauge the relative importance of consumption, leisure, and social capital in terms of utility value, we consider the elasticity of utility with respect to each of these arguments. They indicate that utility is most sensitive to changes in leisure and least responsive to changes in social capital. It is also interesting to note the temporal pattern in these elasticities. As the individuals age, their welfare becomes more responsive to changes in their level of social capital and consumption. In contrast, they become less responsive to changes in leisure. This finding is further support of growing state dependence in preferences.

In our dynamic model, social capital stock accumulation increases the expected cost of engaging in crime, making the occurrence of crime less likely. This life-cycle model of behavior is consistent with the pattern of criminal behavior observed in the age-crime profile. It is interesting to compare the temporal pattern of the age-crime profile of the cohort to which our sample belongs, with the profile of marginal utility of social capital for the sample. Figure 6.2 shows a strong inverse relationship between the two profiles. Our results provide evidence of growing state dependence and reinforcement in non-deviant behavior, and hence increasing costs of deviant behavior, during a period of decline in participation in crime. This suggests that our model provides a possible explanation for the empirical phenomenon of the age-crime profile.

Our model performs well at explaining the decline in participation in crime for the average of our sample. However, the more important question may be how well it explains the behavior of those most at risk of criminality. Our index of social capital stock inherited from the family allows us to investigate this issue. As in Section 6.3, we partition the sample into quartiles on the basis of initial period social capital stock and compare the temporal pattern in the marginal utility of social capital for the first and fourth quartiles, representing the individuals most and least at risk of adult arrest respectively. Figure 6.3 shows that the marginal utility of social capital for individuals in the fourth quartile (low risk group) increases

[25] See Ryder and Heal (1973) and Becker and Murphy (1988).

Figure 6.2. ***The marginal utility of social capital versus the age crime profile***

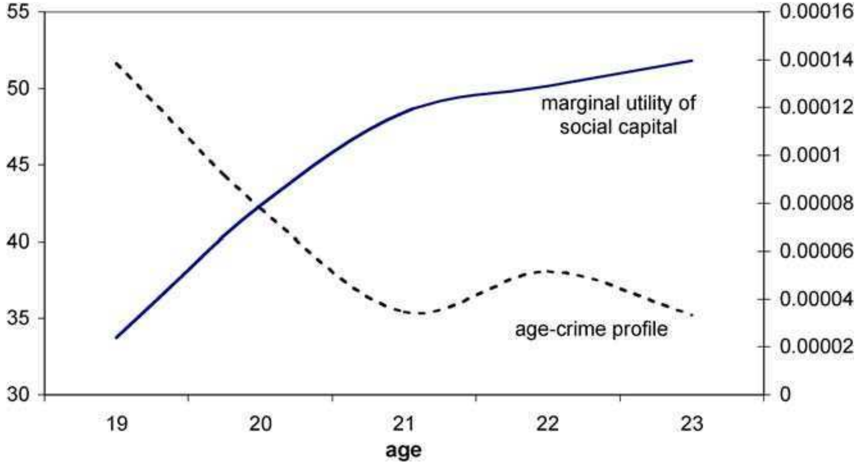
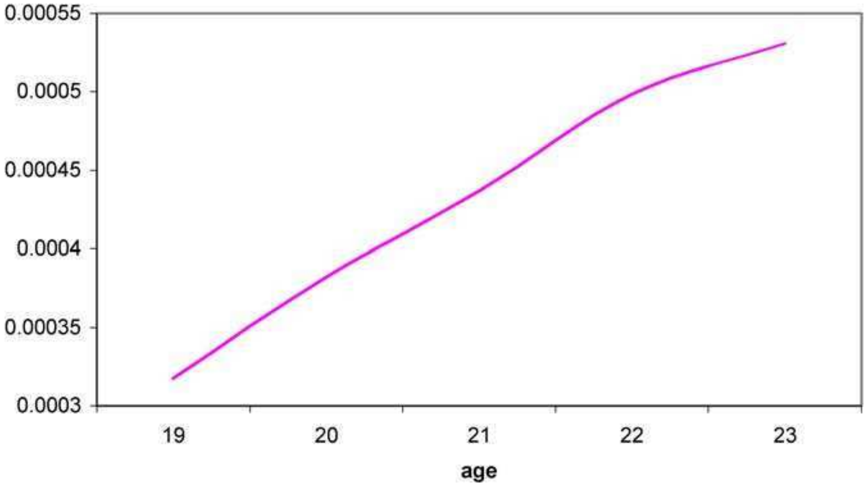
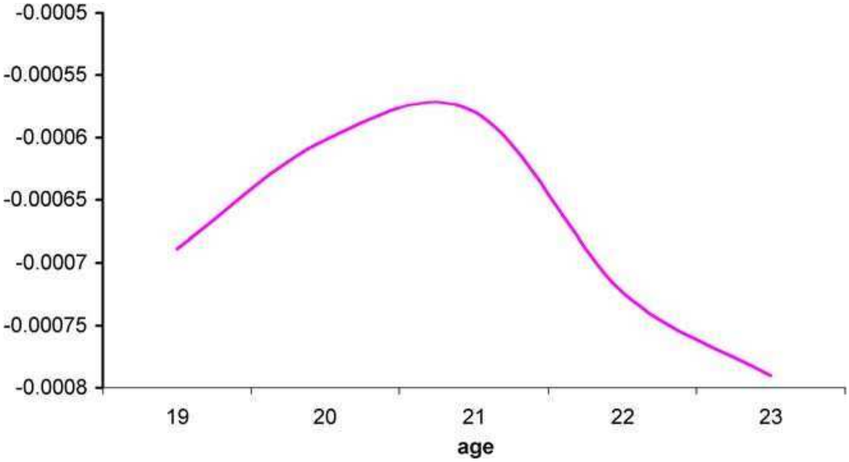


Figure 6.3. ***The marginal utility of social capital for the fourth quartile***



over time, just as it does for the whole sample. The marginal utility of social capital for individuals from the first quartile (high risk group) displays a markedly different temporal pattern, as shown in Figure 6.4. While the value of an incremental increase in social capital increases over the ages 19 to 21, it falls thereafter. Also, the marginal utility of social capital is always negative for this group. The latter finding may be an artifact of

Figure 6.4. The marginal utility of social capital for the first quartile



the assumed functional form for utility. Alternatively, it may be revealing something of a more behavioral nature.

Recall from our earlier discussion involving comparisons among the first and fourth quartiles of arrests and time in crime that we find individuals from the first quartile to be far more likely to be arrested for an income producing crime in any year than those in the fourth quartile. These men appear to be embedded in a criminal peer group by the age of 18, when our study begins, and may consider social capital to hinder their advancement in the criminal peer group. This interpretation is consistent with a negative marginal utility associated with social capital. While state dependence in crime appears to diminish over the ages of 19 to 21, as indicated by the marginal utility of social capital becoming less negative, it strengthens thereafter. This could be evidence of the difficulty these individuals have overcoming the state dependence in criminal culture and successfully building stock in legitimate society. The implication of this is that differences in the level of social capital inherited from the family may explain why some individuals become career criminals, while others experience relatively short careers in crime.

6.5. Conclusion

In this paper we integrate the intertemporal choice and economics of crime literature to develop a dynamic model of criminal choice that focuses on the role of stigma as a deterrent to crime. Current period decisions affect future outcomes by a process of social capital accumulation. Our model assumes that social capital provides a flow of services associated with a good reputation and social acceptance, and that stigmatism associated with arrest reduces an individual social capital stock. In this way we account for the influence of social norms on the decision to participate in crime.

Using data from the 1958 Philadelphia Birth Cohort Study, we find significant empirical support for the dynamic model of crime. The selectivity corrected earnings equation estimates for labor market activities indicate that legal wages are increasing in both human and social capital. Application of a method of simulated moments estimator to the system of Euler equations reveals significant state dependence in preferences, as measured by the stock of social capital. We find that the marginal utility of past investment in social capital is increasing in current investment, implying adjacent complementarity. This leads to growing state dependence over the life-course. Growing state dependence in non-deviant behavior raises the potential cost of engaging in crime, making its occurrence less likely. Therefore, the model provides an explanation of the empirical relationship between aggregate arrests and age.

We also investigate the performance of the model in explaining the behavior of individuals who differ in their degree of being at risk of becoming criminals. Our findings suggest that low levels of social capital inherited from the family may explain why some individuals become career criminals, while individuals who are more richly endowed experience relatively short careers in crime. Also evident from our results is the dynamic nature of the process of criminal choice. The late teenage years to early twenties is a crucial time for making the transition out of crime, even for those most disadvantaged in terms of inherited social capital stock.

This last finding is of particular interest as it raises the issue of preventative policy for youth. While the traditional economic model of crime provides a basis for formulating deterrence policy, it is silent on preventative policy. The debate over whether prison pays indicates that justifying the costs of incarceration at current levels is questionable and that crime prevention policies for crime prone groups are likely to be more attractive on a cost benefit basis (Freeman, 1996). In order to contribute to the policy discussion on preventative policy, however, economics must explore dynamic models of crime that provide a mechanism for understanding the way in which preventative policy impacts individuals' potential criminal behavior. Our results suggest that further development of social capital models of crime to include human capital accumulation may prove to be a fruitful means for exploring this issue.

Acknowledgements

The authors would like to thank participants of the 8th World Congress of the Econometric Society, The Seventh International Conference on Panel Data and seminar participants at Texas A&M University, Rice University, the Second Annual Texas Econometrics Camp, Australian National University, and The University of Melbourne for their insightful comments and criticisms. Special thanks are extended to Peter Hartley and Herb Emery for their valuable and critical suggestions. The usual caveat applies.

References

Akerlof, G.A., Yellen, J.L. (1994), "Gang behavior, law enforcement and community value", in: Aaron, H.J., Mann, T.E., Taylor, T., editors, *Values and Public Policy*, Brookings Institute, Washington, DC, pp. 173–209.

Akerlof, G.A. (1997), "Social distance and social decisions", *Econometrica*, Vol. 65, pp. 1005–1027.

Akerlof, G.A. (1998), "Men without children", *The Economic Journal*, Vol. 108, pp. 287–309.

Altonji, J.G. (1986), "Intertemporal substitution in labor supply: evidence from micro data", *Journal of Political Economy*, Vol. 94, pp. S176–S213.

Beccaria, C. (1963 [1764]), *On Crimes and Punishments*, Bobbs-Merril, Indianapolis.

Becker, G. (1968), "Crime and punishment: an economic approach", *Journal of Political Economy*, Vol. 76, pp. 169–217.

Becker, G. (1991), *A Treatise on the Family*, Harvard University Press, Cambridge, MA.

Becker, G., Murphy, K. (1988), "A theory of rational addiction", *Journal of Political Economy*, Vol. 96, pp. 675–699.

Bentham, J. (1970 [1789]), *An Introduction to the Principles of Morals and Legislation*, Athlone Press, London.

Bowles, S., Gintis, H. (2002), "Social capital and community governance", *The Economic Journal*, Vol. 112 (483), pp. 419–436.

Case, A., Katz, L. (1991), "The company you keep: the effects of family and neighborhood on disadvantaged youths", Working Paper, National Bureau of Economic Research, Cambridge, MA.

Coleman, J.S. (1988), "Social capital in the creation of human capital", *American Journal of Sociology*, Vol. 94, pp. S95–S120.

Coleman, J. (1990), *Foundations of Social Theory*, Harvard University Press, Cambridge.

Durlauf, S.N. (2002), "On the empirics of social capital", *The Economic Journal*, Vol. 112 (483), pp. 459–479.

Ehrlich, I. (1973), "Participation in illegitimate activities: a theoretical and empirical investigation", *Journal of Political Economy*, Vol. 81 (3), pp. 521–565.

Figlio, R. (1994), "Self-reported and officially defined offenses in the 1958 Philadelphia birth cohort", *NATO ASI, Series D, Behavioral and Social Sciences*, Vol. 76, pp. 267–280.

Figlio, M., Tracy, P.E., Wolfgang, M.E. (1991), “Delinquency in a birth cohort II: Philadelphia 1958–1986”, Computer File, Second Release, Philadelphia, PA: Sellin Center for Studies in Criminology and Criminal Law and National Analysts, Division of Booz-Allen and Hamilton, Inc.: Producers, 1990. ICPSR, Ann Arbor, MI.

Freeman, R.B. (1992), “Crime and the economic status of disadvantaged young men”, in: Peterson, G.E., Vroman, W., editors, *Urban Labor Markets and Job Opportunities*, Urban Institute, Washington, DC.

Freeman, R.B. (1996), “Why do so many young American men commit crimes and what might we do about it”, *Journal of Economic Perspectives*, Vol. 10, pp. 25–42.

Ghez, G., Becker, G. (1975), *The Allocation of Time and Goods over the Life Cycle*, Columbia University Press (for N.B.E.R.), New York.

Glaeser, E.L., Sacerdote, B., Scheinkman, J.A. (1996), “Crime and social interactions”, *The Quarterly Review of Economics*, Vol. 111 (2), pp. 507–548.

Glaeser, E.L., Laibson, D., Sacerdote, B. (2002), “An economic approach to social capital”, *The Economic Journal*, Vol. 112 (483), pp. 437–458.

Gottfredson, M.R., Hirschi, T. (1990), *A General Theory of Crime*, Stanford University Press, Stanford, CA.

Grogger, J. (1995), “The effects of arrest on the employment and earning of young men”, *Quarterly Journal of Economics*, Vol. 110, pp. 51–72.

Grogger, J. (1998), “Market wages and youth crime”, *Journal of Labor Economics*, Vol. 16, pp. 756–791.

Hartley, P.R. (1996), “Numerical approximation of a value function in the presence of inequality constraints: an application to the demand for credit cards”, *Journal of Economic Dynamics and Control*, Vol. 20, pp. 63–92.

Heckman, J. (1974), “Shadow price, market wages and labor supply”, *Econometrica*, Vol. 42, pp. 679–694.

Heckman, J. (1979), “Sample selection bias as a specification error”, *Econometrica*, Vol. 47, pp. 153–161.

Hotz, J.V., Kydland, F.E., Sedlacek, G.L. (1988), “Intertemporal preferences and labor supply”, *Econometrica*, Vol. 56, pp. 335–360.

Imai, S., Krishna, K. (2001), “Employment, dynamic deterrence and crime”, Working Paper, National Bureau of Economic Research, Cambridge, MA.

Kim, J., Mueller, C.W. (1978), *Factor Analysis: Statistical Methods and Practical Issues*, Sage Publications, Beverly Hills, CA.

Knack, S., Keefer, P. (1997), “Does social capital have an economic payoff? A cross-country investigation”, *Quarterly Journal of Economics*, Vol. 112, pp. 1251–1288.

La Porta, R., Silanes, F.L., Shleifer, A., Vishny, R.W. (1997), “Trust in large organizations”, *American Economics Association Papers and Proceedings*, Vol. 87, pp. 333–338.

Laub, J.H., Sampson, R.J. (1993), “Turning points in the life course: why change matters to the study of crime”, *Criminology*, Vol. 31, pp. 301–325.

Lochner, L. (1999). “Education, work, and crime: theory and evidence”, University of Rochester Working Paper, No. 465.

Lott, J. (1990), “The effect of conviction on the legitimate income of criminals”, *Economic Letters*, Vol. 34, pp. 381–385.

Maasoumi, E. (1986), “The measurement and decomposition of multi-dimensional inequality”, *Econometrica*, Vol. 54, pp. 991–997.

Maasoumi, E. (1993), “A compendium to information theory in economics and econometrics”, *Econometric Reviews*, Vol. 12, pp. 137–181.

McFadden, D. (1989), "A method of simulated moments for estimation of discrete response models without numerical integration", *Econometrica,* Vol. 57, pp. 995–1026.

McFadden, D., Ruud, P. (1994), "Estimation by simulation", *Review of Economics and Statistics,* Vol. 76, pp. 591–608.

Pakes, A., Pollard, D. (1989), "Simulation and the asymptotics of optimization estimates", *Econometrica,* Vol. 57, pp. 1027–1058.

Putnam, R.D. (1993), *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton University Press, Princeton.

Putnam, R.D. (2000), *Bowling Alone: The Collapse and Revival of Civil of American Community*, Simon Schuster, New York.

Rasmusen, E. (1996), "Stigma and self-fulfilling expectations of criminality", *Journal of Law and Economics,* Vol. 39, pp. 519–543.

Ryder, H.E. Jr., Heal, G.M. (1973), "Optimum growth with intertemporally dependent preferences", *Review of Economic Studies,* Vol. 40, pp. 1–33.

Sah, R.K. (1991), "Social osmosis and patterns of crime", *Journal of Political Economy,* Vol. 99, pp. 1272–1295.

Sampson, R.J., Laub, J.H. (1992), "Crime and deviance in the life course", *Annual Review of Sociology,* Vol. 18, pp. 63–84.

Sellin, T., Wolfgang, M.E. (1964), *The Measurement of Delinquency*, John Wiley and Sons, New York.

Sickles, R.C., Taubman, P. (1997), "Mortality and morbidity among adults and the elderly", in: Rosenzweig, M.R., Stark, O., editors, *Handbook of Population and Family Economics*, North-Holland, Amsterdam, pp. 559–643.

Sickles, R.C., Yazbeck, A. (1998), "On the dynamics of demand for leisure and the production of health", *Journal of Business and Economic Statistics,* Vol. 16, pp. 187–197.

Snyder, H. (2000), *Juvenile Arrests 1999*, Office of Juvenile Justice and Deliquency Prevention, Washington, DC.

Stokey, N.L., Lucas, R.E., Sickles, R.C. (1989), *Recursive Methods in Economic Dynamics*, Haward University Press, Cambridge, MA.

Thurow, L.C. (1969), "The optimum lifetime distribution of consumption expenditures", *American Economic Review,* Vol. 59, pp. 324–330.

Vella, F. (1998), "Estimating models with sample selection bias: a survey", *Journal of Human Resources,* Vol. 33, pp. 127–172.

Visher, C.A., Roth, J.A. (1986), "Participation in criminal careers", in: Blumstein, A., editor, *Criminal Careers and 'Career Criminals', Vol. 1,* National Academy Press, Washington, DC, pp. 211–291.

Waldfogel, J. (1994), "Does conviction have a persistent effect on income and employment?", *International Review of Law and Economics,* Vol. 14, pp. 103–119.

Williams, J., Sickles, R.C. (2002), "An analysis of the crime as work model: evidence from the 1958 Philadelphia birth cohort study", *Journal of Human Resources,* Vol. 37, pp. 479–509.

This page intentionally left blank

Panel Data Econometrics
B.H. Baltagi (Editor)
© 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S0573-8555(06)74007-X

CHAPTER 7

Swedish Liquor Consumption: New Evidence on Taste Change

Badi H. Baltagi[a] and James M. Griffin[b]

[a]Department of Economics, Texas A&M University, College Station, TX 77843-4228, USA
E-mail address: badi@econmail.tamu.edu
[b]Bush School of Government and Public Service, Texas A&M University, College Station, TX, 77843-4220, USA
E-mail address: jgriffin@bushschool.tamu.edu

Abstract

Sweden, like other Nordic Countries, has experienced a dramatic reduction in per-capita liquor consumption that cannot be explained by increased consumption of other alcoholic beverages. Using a panel of 21 Swedish counties and annual data for the period 1956–1999, we estimate that at least 4 structural breaks, representing taste change are necessary to account for this sharp decline in consumption. The first structural break coincides with the 1980 advertising ban, but subsequent breaks do not appear to be linked to particular policy initiatives. Rather, our interpretation of these results is that there is an increasing concern with health issues and drinking mores have changed.

Keywords: liquor consumption, structural break, panel data, taste change

JEL classifications: C23, D12, I10

7.1. Introduction

Alcoholism has historically been viewed as a serious public policy problem in Sweden even though per capita Swedish alcohol consumption is estimated to be the lowest in the E.U.[1] Policy makers have experimented with a variety of policy instruments to discourage consumption dating back to 1865 with bans on personal production and municipal distribution of alcohol. Over the period 1917 to 1955, Sweden even adopted a rationing

[1] See Leifman (2000).

system using non-transferable coupons.[2] Then in 1955, the decision was made to rely upon very high "sin taxes" to discourage consumption.[3] Today, for example, the Swedish tax alone on a fifth of 80 proof whiskey is about $25 compared to $6 per bottle tax in the U.S. Paradoxically, after all these efforts, per capita liquor sales remained relatively high throughout the 1960's and 1970's. Beginning in 1980, per capita sales has declined by about 65% for no apparent policy explanation. A change of this magnitude is huge, prompting a variety of hypotheses. Can such a decrease be explained by traditional price and income effects? Could increases in non-recorded consumption such as private production, cross-border shopping and smuggling account for the decline in official measured sales data? Alternatively, if reductions in measured sales imply real reductions in consumption, could taste change explain such a precipitous decline?

While the path-breaking work of Becker and Murphy (1988) on "rational addiction" has spawned great interest in whether consumers are "rationally addicted" to alcohol (Bentzen *et al.*, 1999), it is unlikely to explain such a dramatic reduction in sales. We believe a more fundamental question is in order for Sweden and possibly other countries: "Are taste changes responsible for such a large and apparently permanent reduction in liquor sales?" To the extent that tastes have changed, can they be linked to health concerns or to changing social mores, or possibly to social policies that discourage alcohol consumption? David Brook's recent book, *BOBOS in Paradise*, argues that the current generation of young, highly successful people have very different attitudes about drinking and health than previous generations. If Brook's conjectures are true, consumers' indifference curves are shifting.

While economic theory posits the importance of tastes, most econometric specifications posit constant preferences or tastes, leaving price and income effects as sole determinants of consumption. Our analysis shows that traditional specifications using price and income effects cannot explain this precipitous decline.[4] Likewise, we show that increased non-recorded consumption, while important, cannot account for such a large decrease in the official sales data.

This study of the determinants of liquor sales in Sweden is distinctive in that its objective is to test for and measure the extent of taste change.

[2] Sweden had a complex rationing system where citizens committees determined how much spirits each adult member of the community could purchase. These decisions were based on such factors as age, family and social responsibilities, and reputation, see Norström (1987).

[3] For a summary of Swedish policy, see Franberg (1987).

[4] Sales data are reported on a 100% alcohol basis.

Previous studies relying on aggregate time-series data have been unable to capture taste change because it is inherently unobservable. We utilize a panel data set spanning the period 1956–1999 for 21 counties in Sweden.[5] This provides a much richer data set than previous aggregate time series studies enabling the capture of taste change. While beer and wine sales are not the focus of this paper, we test whether negative taste changes affecting liquor are offset by positive taste substitution for wine and beer. We also seek to determine whether taste changes are primarily autonomous in nature or can they be linked to the changing age distribution of the population? As Sweden's drinking age population has aged, can the decline in per capita sales be explained by demographics?

Section 7.2 describes past trends in alcohol sales in Sweden as well as the results of past studies. Section 7.3 presents the standard habits persistence model we adopt here to model alcohol sales and discusses the choice of panel data estimators. Section 7.4 presents the econometric results and simulates the role of tastes and price effects to explain alcohol sales in Sweden. Section 7.5 considers the factors that may be producing autonomous taste change. Section 7.6 summarizes the key conclusions.

7.2. Past trends and research findings

Previous research typically focuses on the effectiveness of liquor taxation as a policy instrument and the price elasticity of demand. Of course, given Sweden's penchant for high alcohol taxes, this is a particularly important question for Sweden. While published estimates vary widely across countries, there is general agreement that high prices deter consumption. In their review chapter, Cook and Moore (1999) conclude that economists' most important contribution to this literature is the repeated demonstration that "… consumers drink less ethanol (and have fewer alcohol-related problems) when alcohol prices are increased." Because of the public policy concern over alcoholism, there has been a number of studies examining Swedish alcohol consumption, and in particular liquor consumption because of its primary contribution to total alcohol consumption. The earliest study, by S. Malmquist (1953) examined the period 1923–1939 when non-transferable ration coupon books placed quantitative restrictions on consumption. Even with quantitative limits, Malmquist found a price elasticity of -0.3 and an income elasticity of $+0.3$, suggesting that quantitative restrictions were not entirely binding on all consumers, as evidenced by the statistically significant price and income effects. Subsequently, Sundström

[5] This is annual sales data reported in The National Alcohol Board's, *Alcohol Statistics*.

and Ekström (1962) restricted their sample to the coupon era (1931–1954) and found a similar price elasticity of −0.3 but a higher income elasticity of +0.9. Not surprisingly, when Huitfeldt and Jorner (1982) estimated a demand equation for liquor in the post-coupon era using data for 1956–1968, they found a surprisingly large price elasticity of −1.2 and a 0.4 income elasticity.

More recently, there have been other studies utilizing various demand systems to model own price elasticities for liquor, beer, and wine (Clements and Johnson, 1983; Clements and Selvanathan, 1987; Selvanathan, 1991; Assarsson, 1991; Berggren, 1997). The own price elasticities range from −0.22 to −1.03 while the income elasticities range from +0.6 to +1.5. The patterns of substitution between liquor, beer, and wine were not always plausible, owing probably to the limited relative price variation over the period. While modeling liquor demand as part of an alcohol composite with certain cross equation constraints is theoretically appealing, these systems are not well adapted for estimating dynamic relationships. Neither are they suitable for modeling taste change – the primary focus of our paper.

Another vein of research has focused on the application of rational addiction models to alcohol, see Baltagi and Griffin (2002). Bentzen *et al.* (1999) apply the Becker *et al.* (1994) model to liquor consumption in four Nordic countries, including Sweden. They contrast the standard habits-persistence model (which implies myopic expectations) to the rational addiction model for liquor, wine, and beer. They find “strong” evidence for rational addiction to liquor.

Internationally, the range of elasticity estimates is wide. For example, Cook (1981) surveyed U.S. price elasticity estimates and after finding a range of −2.03 by Niskanen (1962) to 0.08 by Wales (1968) he concluded that “there are no reliable estimates for the price elasticity of demand based on U.S. data.” A survey by Ornstein and Levy (1983) reports a range of −1.0 to −1.5 depending on the country studied. For the U.K., Duffy (1983) uses aggregate quarterly data for the period 1963–1978 and finds a price elasticity of −0.77. Using annual data for the period 1955–1985, Selvanathan (1988) finds a similar elasticity estimate. Using aggregate data covering about 30 years, Clements *et al.* (1997) report results for their estimates of systems of demand equations for Australia, Canada, Finland, New Zealand, Norway, Sweden and the U.K. Their average own-price elasticities are −0.35 for beer, −0.68 for wine, and −0.98 for spirits.

In contrast to aggregate time series studies, panel data studies that bear some similarity to our study include Johnson and Oksanen (1974, 1977) who used Canadian provincial data for the period 1955–1971 and found a price elasticity of −1.77. Baltagi and Griffin (1995) used a panel of U.S.

states for the period 1960–1982, and found a long-run price elasticity of -0.69. The advantage of panel data is the richness of the data set in which consumption patterns vary in different panels over time due to increased variation in price and income. Moreover, it is ideally suited to capture taste changes that are common across the cross section of regions. Aggregate time series are incapable of eliciting taste change because it is unobservable. Cross-sectional data have a similar problem in that tastes are constant at a point in time. Individual survey data such as in Berggren and Sutton (1999) tend to be cross-sectional surveys and thus incapable of measuring inter-temporal taste change.

Interestingly, none of these papers consider the possibility of taste change as an important determinant of consumption. In Cook and Moore's (1999) review chapter, they note that U.S. liquor consumption has also declined significantly since 1980. Likewise, liquor sales has declined sharply since 1980 in Sweden, Norway, and Denmark. Finland experienced a significant decline beginning in the mid-1980's. In sum, anecdotal evidence suggests that the taste change hypothesis may have application beyond Sweden.

Figure 7.1 shows the pattern of sales per capita of liquor, beer, and wine and total alcohol in Sweden for the period 1956 to 1999. Sales are measured on a 100% alcohol basis to facilitate comparison between liquor, beer, and wine. Note that the top line, shows aggregate sales of all forms of alcohol, which has declined by 24% since 1976. Interestingly, virtually all of this decrease can be explained by declining liquor sales. Note also that per capita liquor sales was relatively stable until the late 1970's at approximately 3.8 liters per year. Then liquor sales declined precipitously to 1.3 liters per capita by 1999 – a 65% reduction. In contrast, wine and beer sales show a very different pattern. Per capita wine sales continued to grow steadily over the entire period. Per capita beer sales rose sharply over the period 1965–1975 when Class 2B beer (a 3.5% alcohol beer) was sold in grocery stores. Following abandonment of Class 2B beer, sales per capita declined from 1976 to 1980. Since then per capita beer sales has returned to the levels reached in the 1970's. Clearly, on a 100% alcohol basis, liquor sales has declined to unprecedented levels, while beer and wine have only partially offset this decrease.

One might conclude that the high tax rates on alcohol in Sweden would offer the perfect data set to isolate the effect of price, especially if the decline in liquor sales can be linked to rising liquor prices. But this is not true. When Sweden dismantled its coupon rationing system in 1955, high tax rates on alcohol were already in place. Since then, Swedish authorities have adjusted tax rates upward so that the real prices of liquor, beer, and wine have fluctuated within a modest range of ±15%. Figure 7.2

Figure 7.1. Swedish alcohol sales per capita measured by alcohol content (100%)

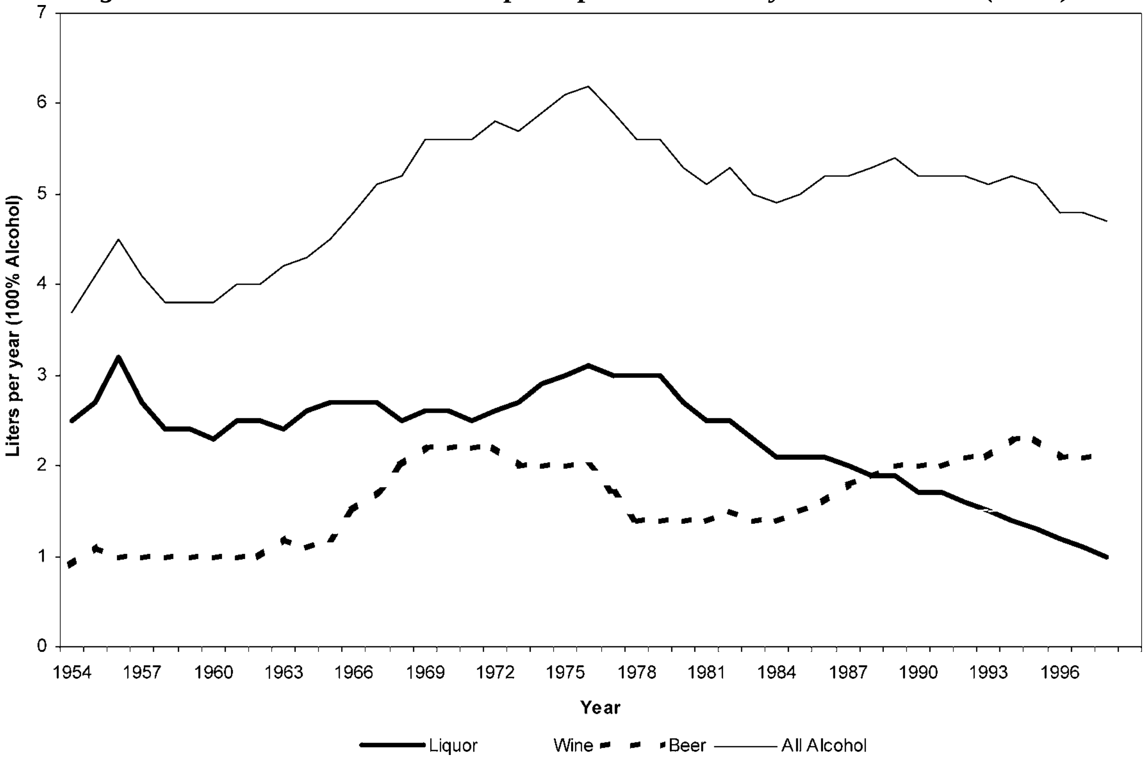
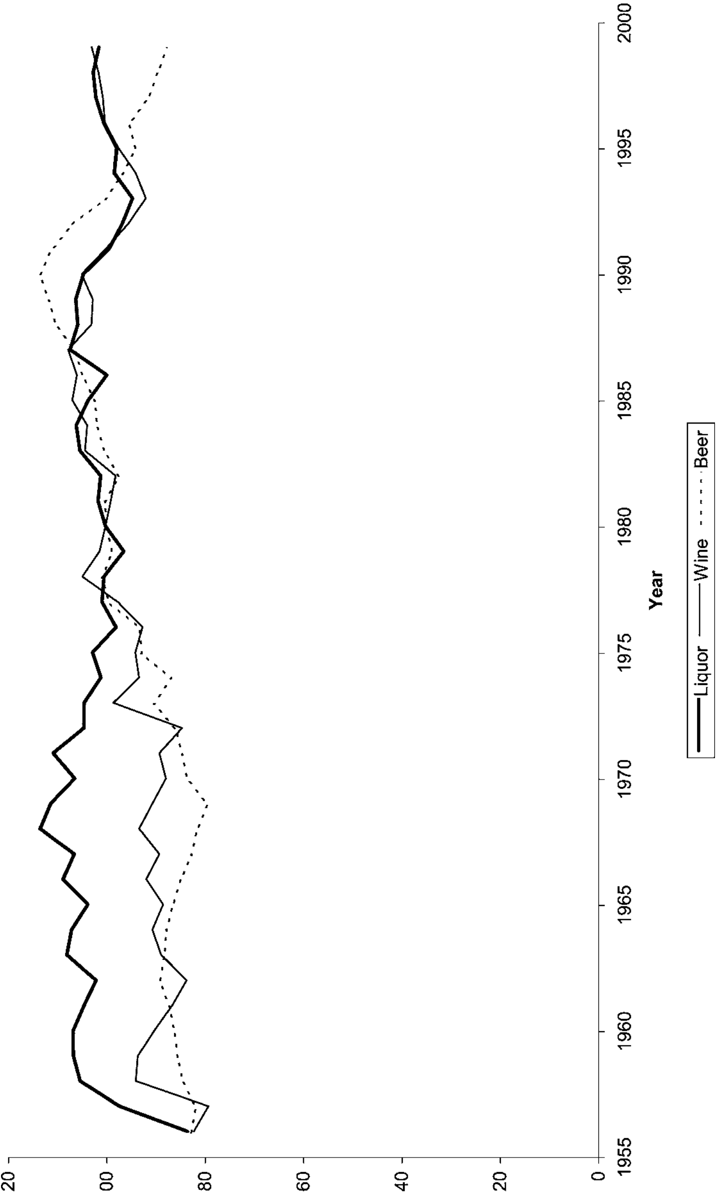


Figure 7.2. Real price of liquor, wine and beer, 1980 = 100



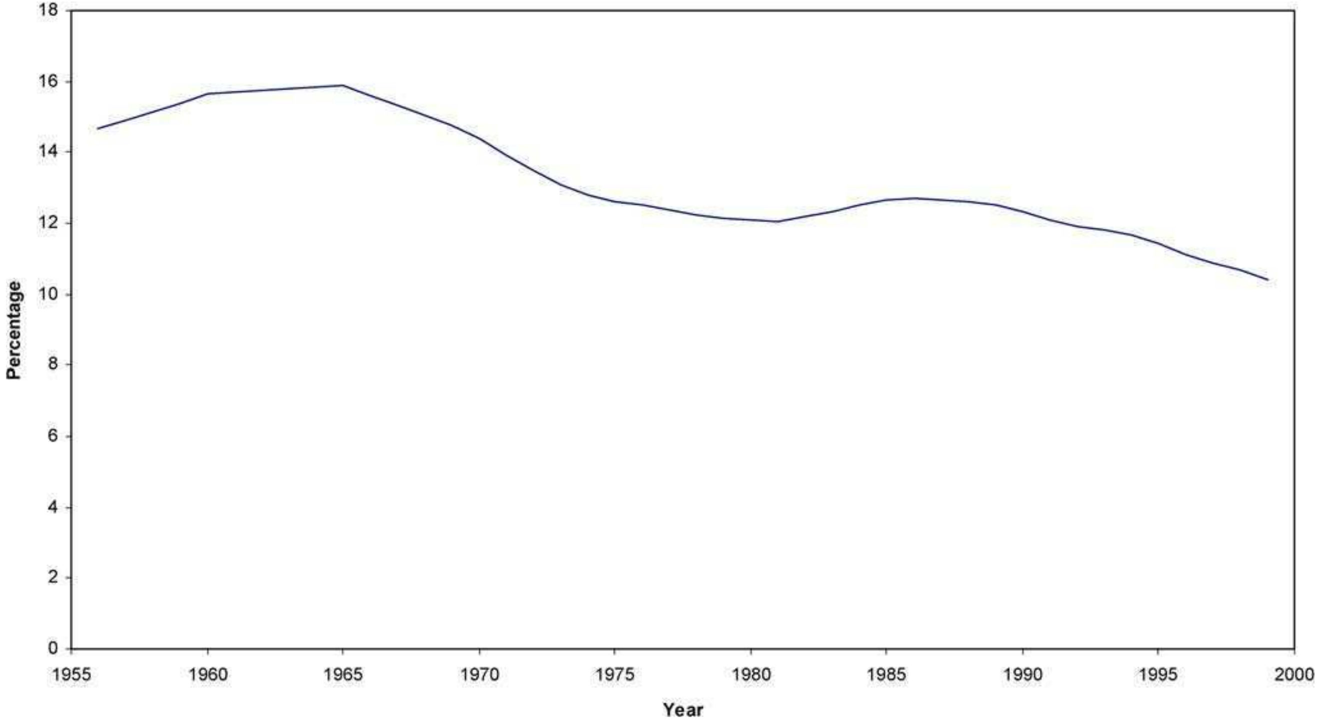
shows the real prices of liquor, beer, and wine for the period 1956–1999. Note that the *relative* price variation between liquor, beer, and wine is even more limited because authorities adjust the tax rates of all three based on the alcohol content of the beverage. Not only is the inter-temporal relative price variation limited, the inter-county variation in the prices of liquor, beer, and wine prices is non-existent. Unlike the U.S., where states impose different taxes creating substantial inter-state price variation, Swedish alcohol prices are uniform across counties, being sold only in state-owned stores. In effect, the inter-county Swedish price history does not allow us to contribute to the issue of price substitutability between liquor, beer, and wine.[6] Instead our focus here is on the substitution between liquor and non-alcohol sales, the relative price variation of which should be adequate to identify this substitution response. This, of course, is critical to the issue of whether sin taxes are an effective deterrent of alcohol sales.

Could the decline in liquor sales be the result of an aging Swedish population? Data from the U.S. National Food Consumption Surveys show that young drinkers, aged 18–24 consume a disproportionally large fraction of beer consumption. It is plausible to expect a similar finding for liquor. Applying this result to alcohol in general, we posit that the age composition of the population serves as a surrogate for taste change. Accordingly, an 18 year old drinker is not equivalent to a 50 year old consumer, both in the amount and type of alcohol consumed. Swedish population data provide detailed population counts by age and by county.[7] There are significant differences both over time and across counties to test the importance of this factor. Figure 7.3 shows the percentage of the population aged 18 to 24. Note that this percentage rose steadily in the post-war period reaching a high of 16% in 1965. As the baby boomers of the post war period aged above 24 and the birth rate continued to decline, the percentage of population aged 18 to 24 has fallen to about 11% in 1999. The obvious question is whether the age composition of the potential drinking-age population can together with price and income explain the decline in liquor sales since 1980. Alternatively, taste change may be largely autonomous in nature, occurring at random times and not directly attributable to some measurable characteristic like the age distribution of the drinking-age public.

[6] Attempts to include the relative prices of liquor, beer, and wine lead to generally insignificant and meaningless results. One should not interpret from this that they are not substitutes, rather the correct inference is that relative price variation between liquor, beer, and wine is insufficient to identify these substitution relationships.

[7] Prior to 1968, censuses were conducted at five year intervals, necessitating interpolation of this variable prior to 1968.

Figure 7.3. Percentage of drinking-population age 18–24, 1956–1999



7.3. The data, model, and choice of panel data estimator

Following previous work,[8] it is reasonable to model liquor demand using a habits persistence model proposed by Houthakker and Taylor (1970). The simplest specification features a log-linear relationship relating per capita liquor sales (C) to the real price of liquor (P), real income per capita (Y) and lagged per capita sales (C_{-1}) as follows:

$$\ln C = \alpha + \beta \ln P + \gamma \ln Y + \lambda \ln C_{-1} + u. \tag{7.1}$$

This specification treats tastes as given, but either in time series over an extended period or a cross section across diverse population groups, it becomes important to include time-varying or cross sectionally varying taste variables (Z_t or Z_i):

$$\ln C = \alpha + \beta \ln P + \gamma \ln Y + \lambda \ln C_{-1} + \Phi Z_t + \Phi^* Z_i + u. \tag{7.2}$$

Cross sectionally varying taste differences, Z_i, which are unobservable, cannot be captured in pure cross section studies, but can be modeled with a panel data set using county specific dummy variables. Likewise, time varying tastes, Z_t, are unobservable, preventing its direct incorporation into purely time series models. But Z_t can usually be estimated as a time dependent dummy variable in a pooled cross section/time series model. Unfortunately, the real price of liquor, P, does not vary over the i counties because prices are uniform at state-run liquor stores.[9] Consequently, individual dummy variables for each time period, Z_t, would be collinear with liquor price, P_t.

Our approach to modeling time-varying tastes are twofold. First, we utilize an explicit age composition variable to reflect taste differences between older versus younger drinkers. Obviously, to the extent that taste changes can be described by differences between the preferences of older versus younger drinkers, it is straightforward to introduce an explicit variable accounting for tastes. Specifically, to capture the effects of differences between older and younger drinkers, we use the percentage of adults 18 to 24 relative to the whole drinking age population, 18 or older.

$$\ln C = \alpha + \beta \ln P + \gamma \ln Y + \lambda \ln C_{-1} + \delta \%\mathrm{AGE18–24} + \Phi^* Z_i + u. \tag{7.3}$$

Presumably, $\delta > 0$ since younger drinkers are likely to drink more intensely.

[8] For example, see Houthakker and Taylor (1970), Johnson and Oksanen (1974, 1977) and Baltagi and Griffin (1995).

[9] Note also that income per capita data in Sweden are available only on a national level.

Second, the other approach to modeling taste change involves testing for structural breaks, particularly over the period 1980–1999, when liquor sales trended downward so sharply. Bai and Perron (1998) develop a methodology for testing for multiple structural breaks, which meets our objective of considering time varying tastes. Our approach is first, using the whole data set to test for a single structural break and then to proceed sequentially testing for subsequent structural breaks as outlined in Bai and Perron. In particular, we consider a partial structural change model where the parameters of price, income and lagged sales are not subject to shifts, but we allow for structural breaks in the time intercepts occurring at unknown dates. This approach is analogous to the time-dependent intercepts, Z_t, in Equation (7.2) except that the structural breaks are less frequent than individual intercepts for each time period and indicate permanent changes for that and subsequent years. Specifically, structural breaks, representing autonomous taste changes, are appended to Equation (7.1) as follows:

$$\ln C = \alpha + \beta \ln P + \gamma \ln Y + \lambda \ln C_{-1} + \theta D_{t-T} + \Phi^* Z_i + u, \quad (7.4)$$

where each structural break, D_{t-T}, spans the period t when it first occurs until T, the end of the sample. In Equation (7.4), θ can be thought of as a vector, reflecting multiple taste changes. Obviously, if there were a statistically significant structural break for each t, then Equation (7.4) would become identical to Equation (7.2). The Bai–Perron procedure enables identification of the most statistically significant changes.

Our preferred estimation approach is the commonly used fixed effects (FE) model incorporating separate intercepts for each county i. Particularly in this case, there is reason to expect taste and other structural differences between counties to be persistent over time. For example, counties in the south of Sweden are close to Denmark, where alcohol prices have traditionally been much cheaper. Furthermore, the north of Sweden has darker winters coupled with a more rural setting – both of which may affect liquor consumption. Thus the fixed effects estimator explicitly enters dummy county variables to reflect differences, Z_i, between counties. This allows for heterogeneity across counties and guards against omitted county-specific variables that may be correlated with the error.

We also employ a fixed effects, two stage least squares estimator (FE-2SLS) to deal with the potential endogeneity of lagged per-capita sales, C_{-1}. Particularly, if the disturbances are autocorrelated, the regression coefficients will be biased. While the FE-2SLS estimator is preferable to the standard FE estimator on purely theoretical grounds, the success of the 2SLS estimator hinges critically on the quality of the instruments, which are typically the lagged values of the price and income variables and possibly the lagged value of the age composition variable. Since price and income do not vary across counties, the estimates for the lagged sales may be very poor. Interestingly, for a dynamic demand model for cigarettes in the U.S., Baltagi *et al*. (2000) show that the out-of-sample performance of the FE-2SLS estimator was inferior to the standard FE estimator.

7.4. Empirical results

7.4.1. Basic habits persistence model with and without age composition

Table 7.1 reports the key results for liquor sales using an explicit measure of taste change, contrasting it to the standard model with no taste change. Rows 1 and 2 utilize a standard habits-persistence equation postulating that all drinking age consumers are homogeneous and that tastes are unchanging over time. Row 1 utilizes the standard fixed effects (FE) estimator, while row 2 utilizes a fixed effects, two stage least squares estimator (FE-2SLS). The results of the basic habits persistence model are quite disappointing. The FE estimator in row 1 indicate a coefficient of 1.03 on lagged sales, indicating such strong habits persistence as to make the implied long run elasticities explosive since $\lambda > 1$. Additionally, the coefficient on price is implausible. The results using FE-2SLS in row 2 are not much better. While the coefficient on the lagged dependent variable falls in the admissible range, price remains with an incorrect sign. Furthermore, the coefficient on income suggests liquor is an inferior good. Both sets of results (rows 1 and 2) suggest a serious specification error. Not surprisingly, price and income are insufficient to describe the precipitous decline in liquor sales since 1980.

Rows 3 and 4 of Table 7.1 introduce taste change due to changing demographics. They include as an explanatory variable, the percentage of the drinking-age population 18 to 24. Note that in both equations, this variable is strongly significant with the correct sign, confirming that younger drinkers consume more liquor than older drinkers. Furthermore, note that the inclusion of % 18–24 causes the coefficients on price and income to become theoretically plausible. The coefficient on the lagged dependent variable, λ exceeds one in row 3, indicating explosive long run responses, but in row 4 (using a FE-2SLS estimator) the coefficient on lagged sales is 0.82. The implied long run price elasticity is -1.34 and the implied long run income elasticity is 0.23.

The long run price elasticity in row 4 indicates a price elastic demand for liquor. To some, this might seem surprisingly large, but we believe it is reasonable. First, it is not entirely outside the range observed in previous studies. As noted above, in Huitfeldt and Jorner's study of Swedish

Table 7.1. Results for basic habits persistence model with and without age composition

								Long-run elasticity		
No.	Estimator	Price	Income	% (18–24)	C_{-1}	R^2	SE	Price	Income	% (18–24)
Liquor sales										
1	FE	0.131	−0.013		1.031	0.959	0.060	–	–	–
		(0.053)	(0.009)		(0.008)					
2	FE-2SLS	0.111	−0.058		0.934	0.953	0.065	1.68	−0.88	–
		(0.058)	(0.022)		(0.042)			(1.71)	(0.39)	
3	FE	−0.146	0.115	0.331	1.025	0.963	0.057	–	–	–
		(0.059)	(0.016)	(0.035)	(0.007)					
4	FE-2SLS	−0.255	0.037	0.406	0.799	0.925	0.081	−1.26	0.18	2.02
		(0.086)	(0.028)	(0.052)	(0.046)			(0.53)	(0.24)	(0.50)
Wine sales										
5	FE	−0.393	0.203	−0.229	0.856	0.992	0.053	−2.73	1.41	−1.59
		(0.037)	(0.035)	(0.029)	(0.015)			(0.73)	(0.25)	(0.31)
6	FE-2SLS	−0.452	−0.059	−0.161	0.983	0.991	0.055	−26.59	−3.47	−9.47
		(0.049)	(0.142)	(0.047)	(0.068)			(78.09)	(32.67)	(22.62)
Beer sales										
7	FE	−0.222	0.651	0.148	0.752	0.920	0.059	−0.90	2.63	0.60
		(0.062)	(0.074)	(0.077)	(0.026)			(0.26)	(0.71)	(0.23)
8	FE-2SLS	−0.217	0.679	0.142	0.739	0.920	0.059	−0.83	2.60	0.54
		(0.065)	(0.133)	(0.081)	(0.056)			(0.83)	(1.13)	(0.31)

liquor consumption, they found long run price elasticities in the −1.2 range. Berggren's recent estimate is −1.03. Second, we believe there is a common sense explanation as well. Liquor is highly taxed in Sweden versus say Denmark and versus the "Duty-Free" shops in the airports. Furthermore, liquor is highly portable, suggesting that there may be considerable leakage in the tax system. Another type of leakage between official sales data and actual consumption is illegal home production of liquor and cross-border purchases. Anecdotal evidence suggests that these sources are substantial. Since smuggling, duty-free, and illegal home production are all likely to be affected by changes in Swedish liquor prices, these high elasticity estimates may reflect substantial leakages as well as actual consumption responses.

Next, we consider the inclusion of taste change driven by demographic factors. The effect of drinking age population appears particularly impressive. The long run elasticity in row 4 of 2.17 coupled with a 20% decline in population aged 18–24, implies that this variable plays a critical role in explaining the secular decline in liquor sales.

Before embracing row 4 of Table 7.1 and the "story" that the decline in liquor sales can be explained by a decline in the percentage of young drinkers, we should look to wine and beer for corroborating evidence. Presumably, if we are observing a taste change induced by a changing age distribution of the population, patterns for liquor would be expected to be operative for wine and beer as well. A lower fraction of younger drinkers, who are presumably heavier drinkers of all forms of alcoholic beverages, should likewise imply negative effects on wine and beer sales. Rows 5 through 8 of Table 7.1 test whether the age composition variable is operative for wine and beer. Note that for wine in rows 5 and 6 of Table 7.1, a higher percentage of younger drinkers leads to *reduced* wine sales. For beer in rows 6 and 7, the coefficient on the % 18–24 is only marginally significant and roughly half the magnitude of the responses for liquor. Intuitively, it seems implausible that a greater fraction of younger drinkers would choose to drink more liquor, less wine, and somewhat more beer. Of course, it may be possible, but a more plausible response is that all forms of alcohol sales would decline with a smaller proportion of young drinkers. Our concern is that the statistical significance of % 18–24 in the liquor demand equation could be spurious. The aging of the Swedish population due to the declining birth rate may simply be spuriously correlated with declining liquor sales.

7.4.2. Tests for autonomous taste change

The competing hypothesis is that taste changes are of an autonomous nature, driven either by changes in attitudes about health and/or policies to discourage drinking. Table 7.2 reports our tests for autonomous taste change by testing for structural breaks. Our procedure for testing for multiple structural breaks is that outlined by Bai and Perron (1998). From Figure 7.1, it is not surprising to note that the most significant structural break occurred for the period 1980–1999. The next most significant break occurred for the period 1993–1999. Next follows statistically significant breaks for the period 1987–1999 and then for the period 1995–1999. Row 1 of Table 7.2 takes the basic habits persistence model and appends structural breaks. Note that with a logarithmic dependent variable, the coefficients for structural changes can be interpreted as short-run percentage reductions in sales with lagged effects entering with time. In order to describe the cumulative effects of the structural breaks, the dynamic structure of the breaks are shown in Figure 7.4, treating the original intercept at one. The cumulative effect of the structural break in 1999 amounts to a 68% reduction in sales for the fixed effects 2SLS estimate. This result confirms that standard price and income effects cannot explain the precipitous decline in liquor sales.

Interestingly, the structural break model avoids the implausible adjustment parameters in the previous model. Row 1 of Table 7.2 also suggests very plausible long-run price and income elasticities – −1.22 for price and 1.25 for income.

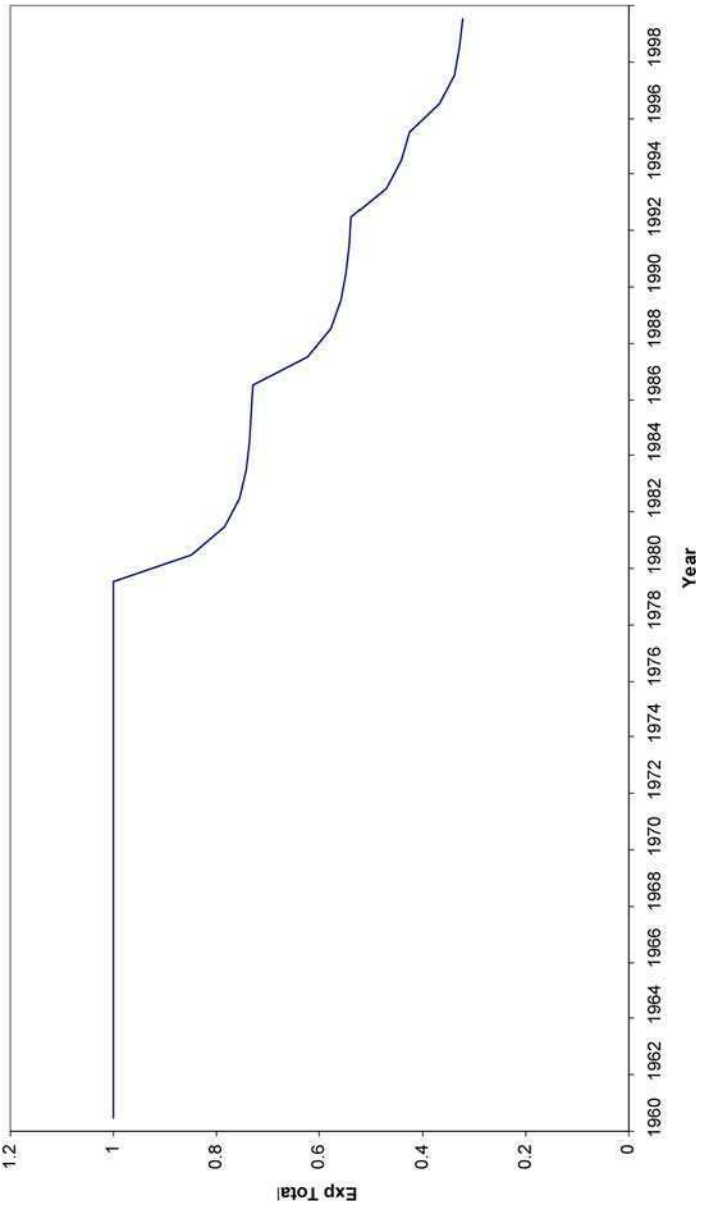
Concerns that the FE estimator may be biased lead us to the FE-2SLS estimator in row 2 of Table 7.2. Note that the coefficient on the lagged dependent variable declines markedly to 0.49, indicating much weaker habits-persistence. While the corresponding short run price elasticity (−0.61) is much larger, the long run elasticity of −1.20 is virtually identical to the long run price elasticity with the FE estimator. Thus one clear picture emerges from either approach to modeling taste changes – the long run price elasticity is price elastic. This implies that price has been a very effective tool in reducing within country alcohol purchases. This does not necessarily translate into effective reductions in measured liquor consumption due to leakages – an issue we return to shortly.

Another related question we seek to answer is whether the negative structural breaks leading to reduced liquor sales have lead to *increased* sales of beer and wine. This is a type of taste substitution hypothesis. Rows 4 through 8 address this question for beer and wine, using the structural breaks identified for liquor. Row 3 for wine sales indicates a short run 3.6% increase in wine sales in 1980, corresponding to the 12.5% reduction in liquor sales. On a pure alcohol basis in 1980, liquor sales accounted for 50% of total alcohol sales with wine contributing 22% and beer 28%. Consequently, a 3.6% increase in wine sales offers only a very small offset compared to the 12.5% liquor sales. Since the time series for strong beer

Table 7.2. Results for autonomous taste change model

											Long-run elasticity	
No.	Estimator	Price	Income	80–99	87–99	93–99	95–99	C_{-1}	R^2	SE	Price	Income
Liquor sales												
1	FE	−0.206	0.212	−0.125	−0.058	−0.054	−0.056	0.831	0.980	0.042	−1.22	1.25
		(0.042)	(0.010)	(0.005)	(0.006)	(0.008)	(0.008)	(0.010)			(0.29)	(0.08)
2	FE-2SLS	−0.613	0.309	−0.163	−0.155	−0.134	−0.138	0.487	0.954	0.064	−1.20	0.60
		(0.072)	(0.017)	(0.008)	(0.012)	(0.014)	(0.014)	(0.032)			(1.14)	(0.06)
Wine sales												
3	FE	−0.520	0.221	0.036	−0.024	−0.027	0.036	0.890	0.992	0.054	−4.73	2.01
		(0.040)	(0.039)	(0.007)	(0.007)	(0.011)	(0.011)	(0.015)			(1.96)	(0.58)
4	FE-2SLS	−0.721	−0.488	0.030	0.008	−0.093	0.093	1.196	0.988	0.065	–	–
		(0.086)	(0.244)	(0.009)	(0.014)	(0.026)	(0.023)	(0.105)				
Beer sales												
5	FE	−1.099	0.893		0.058	−0.089	−0.128	0.767	0.934	0.054	−4.72	3.83
		(0.111)	(0.093)		(0.014)	(0.016)	(0.015)	(0.028)			(2.28)	(1.50)
6	FE-2SLS	−0.761	1.469		0.108	−0.003	−0.130	0.299	0.889	0.070	−1.09	2.10
		(0.167)	(0.187)		(0.022)	(0.030)	(0.020)	(0.121)			(0.38)	(0.43)

Figure 7.4. Measure of autonomous taste change



sales by counties only begins in 1979, it is not meaningful to calculate any offset for increased beer sales. The next most prominent structural break for 1993–1999 appears to offer no evidence of offsetting wine or beer sales as both indicate negative effects that year. The 1987 structural break shows an inconclusive pattern for wine and a positive 5.8% increase for beer sales in the FE model in row 5. But for the period 1995–1999, beer sales suffers an apparent negative structural break in sympathy with liquor, while the coefficients for wine indicate an increase.

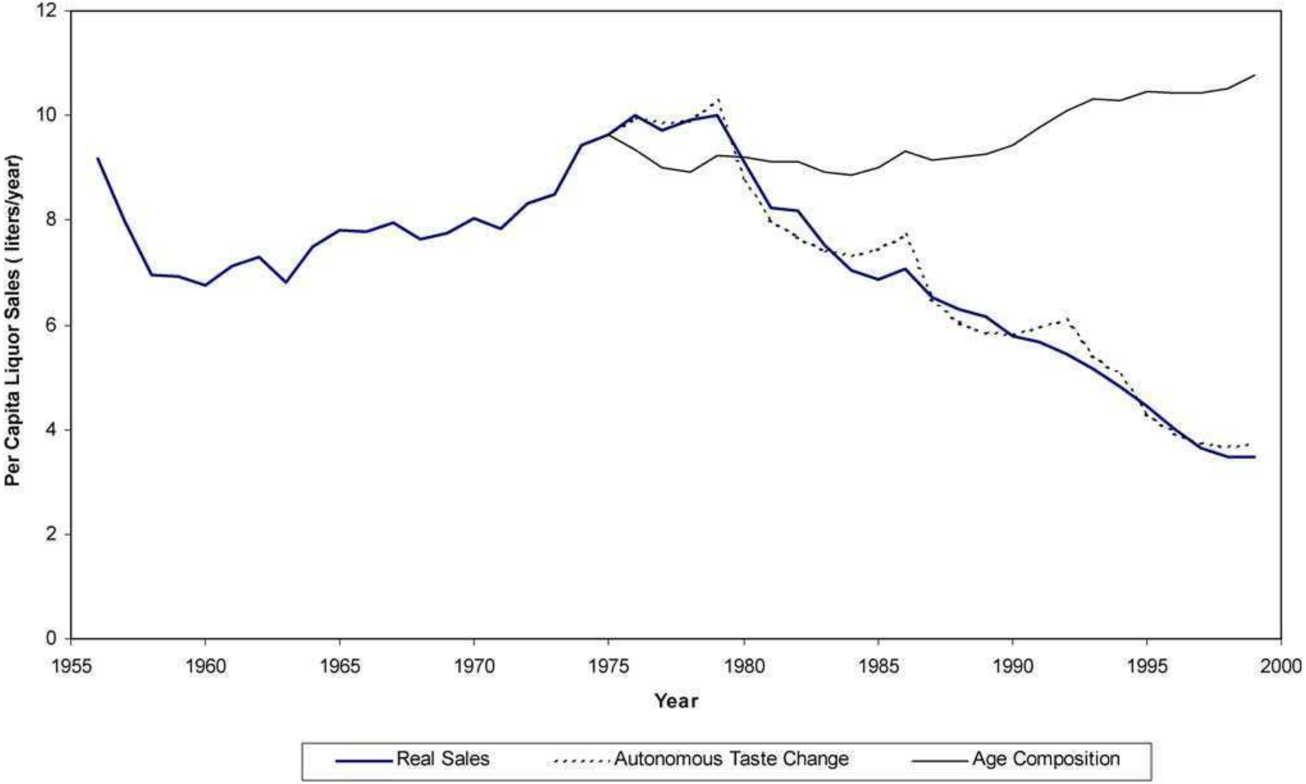
We conclude that the evidence favoring a taste substitution hypothesis from liquor to wine and beer is not particularly convincing. We should expect to see a pattern of positive coefficients on the various structural breaks for wine and beer. This pattern is not supported by the data. Indeed, the cumulative effects on beer are like liquor – negative – while the cumulative effects on wine are ambiguous.

7.4.3. Forecast comparison of two competing types of taste change

In comparing the two alternative hypotheses of taste change, age composition vs. autonomous taste change, it is instructive to compare the forecast performance of the two. Figure 7.5 contrasts the predictive performance of the age-composition taste hypothesis shown in Table 7.1 with the autonomous taste change model described in Table 7.2 featuring structural breaks. Even though the data were estimated from county data, we test the ability of the models to explain aggregate per capita sales for the whole country. We use the parameter estimates in row 4 of Table 7.1 to describe the age-composition hypothesis versus corresponding estimates in row 2 of Table 7.2 to describe the autonomous taste change hypothesis. We perform an in-sample forecast exercise using a *dynamic* simulation beginning in 1975 using forecasted values in period t as actuals for lagged sales per capita in period $t + 1$. Consequently, auto-correlated forecast errors can produce wildly different simulation performance than a series of one-period-ahead forecasts.

As shown in Figure 7.5 the dynamic simulation properties of the two models are quite different. The autonomous taste change model with discrete structural breaks closely tracks the precipitous decline in liquor sales. In contrast, the age composition characterization of taste change fails to explain the decline. We view this as convincing corroborating evidence that while the age distribution variable probably plays some minor role, autonomous taste change is the decisive explanation. Indeed, we found that % 18–24 could be appended to the autonomous taste change model with the former entering as expected and the structural breaks not being materially impacted.

Figure 7.5. Age composition vs. autonomous taste change: forcast performance 1975–1999



7.4.4. Could autonomous technical change really be due to leakages?

Skeptics of the autonomous technical change hypothesis might argue that actual liquor consumption as opposed to official sales data, may not have in reality declined. They would argue that the apparent 72% reduction in measured liquor sales implied by Figure 7.4 could be offset by large increases in non-recorded consumption. Over time, Swedes have become much more mobile with increasing access to lower cost liquor in Denmark and duty-free shopping in airports. Domestic moonshine is another source. Could these factors explain the precipitous decline in domestic liquor sales, so that in reality the Swedes are drinking as much as before 1980?

Kühlhorn and Ramstedt (2000) utilize survey data on total alcohol consumption to argue that non-recorded consumption of total alcohol is both substantial and has increased markedly since 1980. They estimate that as a percent of recorded sales data, non-recorded consumption has increased from 13% in 1980 to 37% in 1997. Leifman's (2000) estimates are substantially lower – 24% in 1995. Even if these estimates are correct and even if non-recorded alcohol consumption is all liquor, no more than half of the apparent autonomous taste change might be attributable to increased percentages of non-recorded consumption.

To test these conjectures more formally, we obtained relative price data for Danish liquor compared to domestic Swedish liquor. The data show that the relative price of Danish to Swedish liquor has remained persistently lower over a period of time dating back to at least 1960. Thus the sharp decline in domestic liquor purchases after 1980 does not appear to be explained by a sharply declining relative price of Danish to Swedish liquor.[10] Likewise, we obtained data on the number of air travel passengers leaving and entering Sweden to measure increased access to duty-free liquor. The data show a persistent growth in international air travel since 1960 – the first year of data. Visually, there is no evidence of a marked acceleration of growth after 1980. While we cannot measure domestic moonshine production, there is no reason to expect a marked increase after 1980. Over this period, the relative price of liquor in Sweden did not appreciably change. One might expect illegal moonshine production to be a function of Swedish liquor prices, but there is no reason to posit a structural increase of massive proportions after 1980.

To formally test the conjecture that increased mobility either to Denmark or international duty-free travel, we regressed the autonomous taste

[10] See Norström (2000) for a post E.U. study of Danish imports into southern Sweden.

change index shown in Figure 7.4 on international passenger travel (leaving and entering Sweden), the relative price of Danish to Swedish liquor, and a time trend. This regression found no statistically significant relation for either the Danish to Swedish liquor price or international air travel. Thus, we conclude that the observed autonomous taste change is real and cannot be simply explained by offsetting increases in non-recorded consumption.[11] At the same time, given the magnitude of the percentage decline in liquor sales implied by Figure 7.4, and the findings of Kühlhorn and Ramstedt (2000) and Leifman (2000), we cannot rule out the possibility that increased non-recorded consumption accounts for some non-trivial portion of the reduction.

7.5. Possible explanations for autonomous taste change

Since autonomous factors, as opposed to leakages, appear to account for the bulk of the precipitous decline in liquor sales since 1980, it is instructive to ask what factors lay behind this huge change shown in Figure 7.4. Could any changes in policy over this period account for such a large change? Alternatively, could such a reduction be due to changing attitudes about liquor?

First, we look to Swedish alcohol policy changes as possible explanations. Interestingly, all forms of alcohol advertising were banned in 1979. These bans included all forms of media advertising and applied to all forms of alcohol. Moreover, these bans have remained in effect since then. Clearly, the major structural break in 1980, corresponds closely with the advertising ban. Interestingly, in 1980, wine sales increased by 3.6% after accounting for price and income effects. One interpretation, is that the advertising ban on alcohol may have induced some substitution to drinks with lower alcohol content. Thus we tend to believe that the advertising ban and the sharp reduction in liquor sales are causally linked even though there was some offsetting increase in wine sales.[12] The difficult question is whether the advertising ban reflects an exogenous or endogenous policy change. On interpretation is that the advertising ban was driven by a changing public's perception of the acceptability of alcohol consumption.

[11] To be clear, these results do not prove that leakages in the form of Danish liquor, duty-free, and moonshine are unimportant. They simply do not explain the autonomous structural break beginning in 1980. Rather, we view our rather price elastic demand response as measuring these factors as well as real reductions in consumption.

[12] Beer sales may also have been favorably affected, but the shortness of the time series for beer prevented accounting for effects in the early years of its sample.

But even if the advertising ban exogenously reduced sales after 1980, its effects should have been realized by the late 1980's, leaving subsequent structural changes unexplained. The advertising ban cannot be the whole story.

Subsequently, the Swedes instituted a number of large public campaigns targeted at the reduction of alcohol consumption particularly by young people. These campaigns have typically taken the form of moral suasion and the links to structural breaks are not particularly strong. In 1986, a program called Bomangruppen was initiated aimed at reducing alcohol consumption among young people. In 1989, yet another program, called Athena, was instituted to reduce alcohol consumption among young people, by imposing harsh penalties on those reselling alcohol to underage drinkers. The 1986 Bomangruppen campaign corresponds to the 1987 structural break. This program seems unlikely to account for the 5.8% reduction in liquor sales in 1987, especially since beer sales increased by a similar percentage that year.

Yet, another potential explanation is the punishment of drunken drivers. Sweden has historically been among the least tolerant European nations for blood alcohol levels. In 1957, the blood alcohol level was reduced from 0.08% to 0.05%. Then in 1990, the blood alcohol limit was further reduced to 0.02% – the lowest in Europe. In contrast, in Italy, the threshold on blood alcohol content is 0.08%. The penalties in Sweden are also quite severe. For a blood alcohol level between 0.03 and 0.10, fines based in part on income are coupled with license suspensions ranging from 2 to 12 month. For blood alcohol levels above 0.1%, drivers lose their license for a minimum of 12 months and face 1 to 2 months in jail.[13]

While one can point to advertising bans and stricter blood alcohol limits as exogenous factors reducing alcohol consumption, they signal a less tolerant public attitude towards drinking, particularly of beverages like liquor with high alcohol content. In many ways, these legal changes may simply reflect the changing attitudes of voters.

David Brooks's (2000) *BOBOS in Paradise* provides an insightful commentary on the current generation of young, highly successful people, who tend to impart their values to the broader population. Brooks characterizes this group as "bourgeois bohemians" (i.e., Bobos in Brooks's vernacular) and proceeds to describe their cultural differences vis-a-vis the same socio-economic group of the previous generation. One of the prominent differences is the changing attitude about drinking. Brooks makes the following incisive comments about Bobos' pleasures:

[13] See *On DWI Laws in Other Countries*, U.S. Department of Transportation, DOT HS 809 037, March 2000.

> The Bobos take a utilitarian view of pleasure. Any sensual pleasure that can be edifying or life-enhancing is celebrated. On the other hand, any pleasure that is counterproductive or dangerous is judged harshly. So exercise is celebrated, but smoking is now considered a worse sin than at least 5 of the 10 commandments. Coffee becomes the beverage of the age because it stimulates mental acuity, while booze is out of favor because it dulls the judgement. Brooks (2000, p. 199)

Brooks goes on to describe how parties have changed:

> Now parties tend to be work parties; a glass or two of white wine, a little networking with editors and agents, and then it's home to the kids. Almost nobody drinks at lunch anymore. People don't gather around kitchen tables staying up nights imbibing and talking. Brooks (2000, p. 201)

While Brooks' lively and entertaining book offers little in the way of hard statistical evidence to support his descriptions, his generalizations strike many accordant strings with our own perceptions of reality both in the U.S. and Europe. Certainly, the empirical evidence here fits directly with Brooks' conjecture. The time period, 1980–2000, matches the time period for the ascendancy of the Bobo mentality. Brooks emphasis on health effects explains why liquor, as opposed to wine and beer, appears to have been singled out as the culprit and became the object of sharply reduced consumption. Likewise, the increased emphasis on driver safety and reduced alcohol limits suggests two things. First, as indicated by their stringency, Sweden, among all European countries, is on the forefront with this attitude. Second, the stricter drinking and driving laws in Sweden are indicative of changing attitudes about safety and health. Following Brooks, these laws, like reduced liquor consumption, are the effect of changing attitudes and not the cause of the reduction in liquor consumption. The most interesting question is why the values of the Bobos have evolved as they have. Brooks does offer some conjectures, but that takes us well beyond the issue at hand. For our purposes here, we lay the immediate cause to modern society's attitudes about pleasures and the new primacy about health.

7.6. *Conclusions*

Our study leads to a number of interesting conclusions. Like previous studies of liquor consumption in Sweden, we affirm that the price elasticity of demand appears highly price elastic compared to the U.S. and other countries. Not all of this price response necessarily implies lower consumption levels, however. Lower taxes in nearby Denmark, duty-free shops for international travel, and illegal moonshine, reflect a partial substitution of these sources for domestic liquor, thereby increasing the observed price elasticity estimate.

Even though price effects are substantial, they cannot explain the unprecedented decline in per capita Swedish liquor sales since 1980. Our best evidence suggests that autonomous taste changes are responsible for this decline. We found evidence of at least 4 structural breaks starting in 1980. Increased leakages, or non-recorded consumption, probably accounts for a portion of this reduction, but there remains a large, apparently real reduction in sales. While the 1980 structural break coincides with the 1979 ban on all advertising of alcohol in Sweden, it cannot explain subsequent structural breaks. Likewise, we found evidence of new, much stricter blood alcohol limits for drivers. We interpret this, like the advertising ban, as an effect of a new much more health-conscious generation with very different drinking mores, rather than the proximate cause.

Finally, at least for liquor sales, we believe similar taste changes may be at work in other industrialized nations. Demand specifications looking only to price and income effects must consider the possibility of an independent role for taste change or a much more elaborate model of consumer choice.[14] Panel data sets of the type employed here seem ideally suited for such investigations.

Acknowledgements

The authors wish to acknowledge the generous financial support of the Bush Program in the Economics of Public Policy and the Texas Advanced Research Program. We also wish to thank Nils Lundrgen, Hank Jenkins-Smith, two anonymous referees and Bertil Vilhemson for access to Swedish data. Sheng Sheu provided excellent research assistance.

References

Assarsson, B. (1991), “Alcohol pricing policy and the demand for alcohol in Sweden 1978–1988”, Working Paper 1991:19, Department of Economics, Uppsala University.

Bai, J., Perron, P. (1998), “Estimating and testing linear models with multiple structural changes”, *Econometrica*, Vol. 66, pp. 47–78.

Baltagi, B.H., Griffin, J.M. (1995), “A dynamic demand model for liquor: the case for pooling”, *The Review of Economics and Statistics*, Vol. 77, pp. 545–553.

[14] Stigler and Becker (1977) have posed the question that what appears as taste change in the simple standard consumer demand model might still be compatible with constant tastes in a more elaborate model of consumer choice. Perhaps it is possible to postulate and to empirically estimate a model in which the “full price” of liquor, including the increased health costs, is sufficient to explain this phenomenon.

Baltagi, B.H., Griffin, J.M. (2002), “Rational addiction to alcohol: panel data analysis of liquor consumption”, *Health Economics,* Vol. 11, pp. 485–491.

Baltagi, B.H., Griffin, J.M., Xiong, W. (2000), “To pool or not to pool: homogeneous versus heterogeneous estimators applied to cigarette demand”, *Review of Economics and Statistics,* Vol. 82, pp. 117–126.

Becker, G.S., Murphy, K.M. (1988), “A theory of rational addiction”, *Journal of Political Economy,* Vol. 96, pp. 675–700.

Becker, G.S., Grossman, M., Murphy, K.M. (1994), “An empirical analysis of cigarette addiction”, *American Economic Review,* Vol. 84, pp. 396–418.

Bentzen, J., Ericksson, T., Smith, V. (1999), “Rational addiction and alcohol consumption: evidence from the Nordic countries”, *Journal of Consumer Policy,* Vol. 22, pp. 257–279.

Berggren, F. (1997), “The demand for alcohol in Sweden 1985–1995, a system approach”, *Studies in Health Economics, Vol. 18,* Department of Community Medicine, Institute of Economic Research, Lund University.

Berggren, F., Sutton, M. (1999), “Are frequency and intensity of participation decision-bearing aspects of consumption? An analysis of drinking behavior”, *Applied Economics,* Vol. 31, pp. 865–874.

Brooks, D. (2000), *BOBOS in Paradise,* Simon & Schuster, New York.

Clements, K.W., Johnson, L.W. (1983), “The demand for beer, wine and spirits: a system-wide analysis”, *Journal of Business,* Vol. 56, pp. 273–304.

Clements, K.W., Selvanathan, E.A. (1987), “Alcohol consumption”, in: Theil, H., Clements, K.W., editors, *Applied Demand Analysis: Results from System-Wide Approaches*, Ballinger Publishing Co, Cambridge, MA, pp. 185–264.

Clements, K.W., Yang, W., Zheng, S.W. (1997), “Is utility addictive? The case of alcohol”, *Applied Economics,* Vol. 29, pp. 1163–1167.

Cook, P.J. (1981), “The effect of liquor taxes on drinking, cirrhosis and auto accidents”, in: Moore, M.H., Gerstein, D., editors, *Alcohol and Public Policy: Beyond the Shadow of Prohibition*, National Academy Press, Washington, DC, pp. 255–285.

Cook, P.J., Moore, J.M. (1999), “Alcohol”, Working Paper No. 6905, National Bureau of Economic Research, Cambridge, MA.

Duffy, M.H. (1983), “The demand for alcohol drink in the United Kingdom, 1963–1978”, *Applied Economics,* Vol. 15, pp. 125–140.

Franberg, P. (1987), “The Swedish snaps: a history of booze, bratt, and bureaucracy: a summary”, *Contemporary Drug Problems,* Vol. 14, pp. 557–611.

Houthakker, H.S., Taylor, L.D. (1970), *Consumer Demand in the United States: Analyses and Projections,* Harvard University Press, Cambridge, MA.

Huitfeldt, B., Jorner, U. (1982), “Demand for alcoholic beverages in Sweden (Efterfrågan påreusdrycker i Sverige, SOU 1982: 91)”, The Temporary Government Committee on Alcohol Policy.

Johnson, J.A., Oksanen, E.H. (1974), “Socio-economic determinants of the consumption of alcoholic beverages”, *Applied Economics,* Vol. 6, pp. 293–301.

Johnson, J.A., Oksanen, E.H. (1977), “Estimation of demand for alcoholic beverages in Canada from pooled time series and cross sections”, *Review of Economics and Statistics,* Vol. 59, pp. 113–118.

Kühlhorn, E., Ramstedt, M. (2000), “The total amount of alcohol consumed in Sweden – recent findings and methodological considerations”, in: Holder, H., editor, *Sweden and the European Union: Changes in National Alcohol Policy and Their Consequences,* Almqvist and Wiksell, Stockholm.

Leifman, H. (2000), "Non-recorded alcoholism in the European union", 2nd Research Conference of the European Comparative Alcohol Study, Stockholm.

Malmquist, S. (1953), "A statistical analysis of the spirits retailing in Sweden", *SOU*, Vol. 52, pp. 160–191.

Niskanen, W.A. (1962), *The Demand for Alcoholic Beverages: An Experiment in Econometric Method*, Rand Corporation, Santa Monica, CA.

Norström, T. (1987), "The abolition of the Swedish alcohol rationing system: effects on consumption distribution and cirrhosis mortality", *British Journal of Addiction*, Vol. 82, pp. 633–642.

Norström, T. (2000), "Cross-border trading of alcohol in southern Sweden – substitution or addition", in: Holder, H., editor, *Sweden and the European Union: Changes in National Alcohol Policy and Their Changes*, Almqvist and Wiksell, Stockholm.

Ornstein, S.I., Levy, D. (1983), "Price and income elasticities of demand for alcoholic beverages", in: Galanter, M., editor, *Recent Developments in Alcoholism, Vol. I*, Plenum Publishing, New York, pp. 303–345.

Selvanathan, E.A. (1988), "Alcohol consumption in the United Kingdom, 1955–1985: a system-wide analysis", *Applied Economics*, Vol. 20, pp. 188–194.

Selvanathan, E.A. (1991), "Cross-country alcohol consumption comparison: an application of the Rotterdam demand system", *Applied Economics*, Vol. 23, pp. 1613–1622.

Stigler, G.J., Becker, G.S. (1977), "De gustibus non est disputadum", *American Economic Review*, Vol. 67, pp. 76–90.

Sundström, Å, Ekström, E. (1962), "The beverage consumption in Sweden (Drydkeskonsumtionen i Sverige)", The Industrial Institute for Economic and Social Research, Stockholm.

Wales, T. (1968), "Distilled spirits and interstate consumption effects", *American Economic Review*, Vol. 58, pp. 853–863.

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74008-1

CHAPTER 8

Import Demand Estimation with Country and Product Effects: Application of Multi-Way Unbalanced Panel Data Models to Lebanese Imports

Rachid Boumahdi[a], Jad Chaaban[b] and Alban Thomas[c]

[a]University of Toulouse, GREMAQ and LIHRE, 21 Allée de Brienne, Bâtiment, F-31000 Toulouse, France
E-mail address: rachid.boumahdi@univ-tlse1.fr
[b]University of Toulouse, INRA-ESR, 21 Allée, de Brienne, F-31000 Toulouse cedex, France
E-mail address: chaaban@toulouse.inra.fr
[c]University of Toulouse, INRA-LERNA, 21 Allée, de Brienne, F-31000 Toulouse cedex, France
E-mail address: thomas@toulouse.inra.fr

Abstract

This paper revisits the issue of estimating import demand elasticities, by considering a variety of unobserved components in international commodity transactions. We use highly disaggregated import data for a single importing country, Lebanon, to estimate a flexible AIDS demand model incorporating a multi-way error component structure for unbalanced panel data. This framework is shown to accommodate for product, country, and time effects as separate unobserved determinants of import demand. Results for major agricultural commodities show that the devised empirical specification is mostly supported by the data, while no correlation exists between import prices and unobserved product or country effects.

Keywords: unbalanced panel data, multi-way error components, trade, AIDS demand models

JEL classifications: C23, D12, F17

8.1. Introduction

Empirical analysis of international trade patterns between regional blocks has emerged over the recent years as a major tool for governments and international organizations. In particular, the current trend toward trade liberalization through dismantlement of tariffs and NTMs (Non Tariff

Measures) implies a need for prediction capabilities from countries and institutions, in order to assess the impact of import price changes in terms of economic growth and welfare (see Panagariya, 2000 for a survey on new developments regarding trade liberalization). A parallel evolution to the WTO rounds has been the design and implementation of bilateral or multilateral trade agreements, generally denoted PTAs (Preferential Trade Agreements). Under such agreements, relative import prices from different regions of the world will be modified from the prospect of a single country, and trade diversion is expected. In other words, reducing or eliminating trade barriers altogether is likely to result in shifts in the relative demand for imports, these shifts being presumably related to the intensity of customs tariffs reduction, but also to the nature of the commodities under consideration. This is also likely to affect the growth potential of developing countries (see Winters, 2004).

International trade economists concerned with empirical evaluations have generally proceeded by considering two approaches: General Equilibrium (GE) models and "Gravity" equations. These approaches suffer from many disadvantages. GE models contain calibrations that are often highly debatable, and involve complex numerous equations that are quite demanding from a data perspective. Gravity based models can only explain aggregate trade creation and diversion, and fail to explain how distant countries have increasing trade among each others. All in all, these models have largely ignored the fact that the sensitivity of demand to import price may be highly dependent on the heterogeneity of the underlying trade sections.

Most applied studies dealing with trade patterns or PTAs integrate the notion of trade or import elasticity, and acknowledge the importance of a consistent and somehow accurate estimate for it. However, from an empirical point of view, such elasticity has in most cases been estimated from time series data only (see Marquez, 1990, 1994). The lack of highly disaggregated trade data is an obvious reason, but also the fact that General Equilibrium or macro-economic models in general do not require sector or commodity-wise import elasticity measures.

The issue of heterogeneous import price elasticities is also particularly important when investigating another trend in trade relations between regions, namely the reduction in export and production subsidies for agricultural products. While the debate over the perverse effects of agricultural subsidies in the US and Europe on the welfare of developing countries is likely to be on the political agenda for many years, the restructuring of European agriculture is also likely to involve some degree of export price harmonization. To this respect, the provision of a consistent estimate for import price elasticity in developing countries at a commodity-group level is certainly helpful.

Given this, there are important differences in the modeling of international trade patterns, depending on whether time series or cross-sectional/panel data are used. In a time series context, which is the most widely used for this type of analysis, the demand for imports originates from a single country, that is, is equivalent to a representative consumer facing a system of relative prices from the rest of the world. Only a limited number of commodities are considered, either at an aggregate level of all imports (macro-economic models) or a series of imported goods (sectoral or individual commodity analysis). In the first case, the breakdown of total imports across sectors (agriculture, industry, services) is not known because of aggregation. In the second case, separability assumptions may be required if no data are available on substitute goods. This means nevertheless that, with time-series data, import decisions are assumed to originate from the country as a whole and vertical integration between importers and final consumers (households, producers, etc.) can be assumed.

When cross-section or panel data are used instead, disaggregated information becomes available at the commodity level. Individual transactions in the form of import/exports data records from customs administration may be used, and in this case, import decisions are then observed from importers and not from the final consumers' perspective. The important consequence is that substitution possibilities between different commodities are likely to be far less important than in the time-series framework, because most importers are specialized in particular products or commodity-groups. On the other hand, substitution patterns for the same product but between competing countries or regions become highly relevant.

The present paper uses, for the first time to our knowledge, individual importer data on a highly disaggregated basis (daily transactions), made available by the Lebanese Customs Administration. The amount of information on import prices and values allows us to consider estimation of a demand system for imports, treating importers as consumers benefiting from competition among different supply sources. To be in line with consumer theory while selecting a flexible form for demand, we adopt the AIDS (Almost Ideal Demand System) specification on import shares, which is consistent with aggregation of individual preferences. We consider a single demand share equation for European agricultural products that are exported to Lebanon, in order to keep the number of estimated parameters to a minimum. The approach used in this paper is to concentrate on a limited number of product categories, namely agricultural products, and to assume that expenditure decisions on these commodities can be separated from other import decisions regarding non-agricultural goods. We further assume separability between agricultural product groups.

While the AIDS model is a suitable choice for demand analysis and statistical inference regarding trade substitution elasticities, it only contains information on price and total expenditure in its basic specification. In order to accommodate for non-price effects that may affect the demand share of commodities exported from some (European) countries, we consider an error components specification that explicitly captures unobserved heterogeneity, both across products and across countries. Since the usual random error-components specification entails linear additive heterogeneity terms, this implies that the latter are in fact interpreted as heterogeneous slopes in the underlying import demand function.[1]

The literature on multi-way error components has developed recently in the direction of multi-way structures, involving balanced or unbalanced panel data models, see Antweiler (2001), Baltagi *et al.* (2001), and Davis (2002). Although random effects and fixed effects constitute commonly used specifications in practice, there have not been many applications dealing with specification choice in the case of multi-way panel data. The three papers cited above propose empirical applications but do not deal with the problem of choosing between a random or a fixed effects model. Moreover, the first two references propose nested error components structures and do not compare with more general specifications of error components. On the other hand, Davis (2002) suggests matrix size-reduction procedures for implementing estimators of more general models with multi-way error components. In this paper, we consider a three-way error components model and discuss several empirical issues: (a) The choice of a nested vs. non-nested error-components specification; (b) The number of error components; and (c) The exogeneity test for right-hand side variables, in the sense of a possible correlation with multiple effects.

The general model we consider is a three-way error component regression with unbalanced panel data, in which product, country and time effects are introduced. This is an extension of a limited series of empirical studies in the literature, that until now have considered only a small number of products (possibly in the form of an aggregate commodity) and country-specific import equations instead of country effects in addition to unobserved heterogeneity related to products (Marquez, 1990, 1994). It should be noted that recent papers have estimated panel data models for import demand using importer and exporter effects (Baltagi *et al.*, 2003; Egger and Pfaffermayr, 2003). These papers however use aggregate data of bilateral trade flows between countries, and do not discuss fixed vs. random effects specifications. The fact that product and country effects may exist and be correlated with import prices is an important empirical issue,

[1] This is because the share of a good in total expenditure is equal to the derivative of the log of expenditure with respect to the price of the good. Hence, the intercept term in any given equation is the slope of the log price term in the (log) expenditure function.

as this would detect the role of unobserved national brand image and/or quality effects in import demand shares (Crozet and Erkel-Rousse, 2004).

This paper makes several contributions to the existing empirical literature. First, we provide a convenient framework for dealing with import demand models with highly disaggregated international trade data, by allowing for country and commodity effects. As mentioned above, the model is particularly useful when unobserved product quality and/or national image are expected to influence the level of regional import demand shares. Second, we show how specification checks can be conducted on the structure of the error term (multi-way, unbalanced panel data) and on the relevance of the random effects assumption. When comparing estimates obtained under the fixed effects vs. the random effects specification, inference can be drawn on the correlation between price levels on the one hand, and unobserved heterogeneity related to country and/or product.

The paper is organized as follows. In Section 8.2, we briefly discuss the properties of import price models, and introduce the notation of our AIDS demand model. In Section 8.3, the estimators for linear panel data models with (non-)nested unbalanced structures are presented. The Lebanese customs data used in the empirical analysis are presented in Section 8.4. Estimation results are in Section 8.5, where we perform a series of specification checks (choice between random-effects and fixed-effects) and test for the presence of these effects. Parameter estimates are then used to compute own- and cross-price elasticities of Lebanese imports between different export regions, allowing one to predict the expected change in import shares when the price of exporting countries is modified. Concluding remarks are in Section 8.6.

8.2. The flexible import model

In the literature on import demand and import price elasticity, trade flows of commodities between countries are often treated as consumer goods whose demand is the result of a utility maximization problem. In this respect, obtaining a consistent model for evaluating price and income effects using trade data should not be much different from the standard applied demand analysis with retail consumer goods. The diversity of products available in export markets can even be made similar to the number of consumers' goods in home stores, by defining aggregate product categories in an adequate way. The advantage with trade flows however, is the fact that the number of supply sources for the same product may be restricted by constructing country-wise or region-dependent export and import quantities and price indexes. By doing so, the researcher implicitly considers some degree of homogeneity in goods exported from a particular country or part of the world. Intuitively, such specific component contained in national or regional commodities, considered from the importing country point of view, will be less difficult to identify if products are observed at a highly disaggregate level. French red wine or Egyptian cotton for instance would be part of the general “Drinks” or “Agricultural products” category, and would thus loose their specific national image.

Although the importance of country-specific components of exported goods in determining import shares may be a rather recent empirical issue, it is nevertheless consistent with the Armington model assumption that products are geographically differentiated (Alston *et al.*, 1990). Final consumers may actually perceive different characteristics of the goods as resulting from national differences, yet the researcher often has to treat such characteristics as unobserved heterogeneity components. In any case, whether the latter is labeled “quality”, “innovation” or “national image” in the literature, these components can often be assumed as originating from a diversification strategy implemented by exporters. Crozet and Erkel-Rousse (2004), for instance, propose an Armington model of trade, that is, a multi-country model of imperfect competition with heterogeneous products. They assume country-specific quality weights to enter products’ sub-utility functions of a representative consumer, to end up with a log-linear representation of import shares as a function of relative prices and relative quality scores. They claim that traditional trade models generally ignore the dimension of product quality, leading to excessively low trade price elasticities. Their model estimation, using survey data as proxies for unobserved quality for European products, reveals that controlling for “observed” quality results in higher own-price elasticities of imports.

In the literature, most models of import price elasticity do not account for quality as such, and make strong assumptions of either block separability or product aggregation to analyze trade on very specific commodities or bilateral trade patterns. The problem with product aggregation is that only country-specific quality components may be identified, as all products from a single export source are considered the same (see Hayes *et al.*, 1990). This implies that perfect substitutability exists among goods within a single commodity group. As far as block separability is concerned, this restriction allows for source differentiation across countries and/or regions, for very specific products (see Satyanarayana *et al.*, 1997). However, it is not always consistent with the theory of consumer demand regarding preference aggregation or expenditure homogeneity (Panagariya *et al.*, 1996). Moreover, substitution patterns are necessarily limited in such a framework and may not be very realistic. As pointed out by Andayani and Tilley (1997), block separability and product aggregation are rather strong assumptions to make, and a more robust way of proceeding with empirical analysis is to consider a more general model of demand.

Modeling expenditures in international economics would therefore require the specification of a demand system that would satisfy basic economic assumptions on consumer (importer) behavior. This demand system should also be simple enough to estimate, and its nature should allow for direct and straightforward inference on consumer reaction to prices. At the same time, an important requisite that has gained much attention in the past decade is that such a demand system should be consistent with aggregation. In other words, the final demand for a given good on a market should be obtained by direct aggregation of individual consumer demands. We follow the approach of Andayani and Tilley (1997) by adopting the Almost Ideal Demand System (AIDS) developed by Deaton and Muellbauer (1980). The rest of the section is devoted to a brief exposition of this model, in particular its features when adapted to trade issues.

Let the subscripts i and j denote distinct imported goods, and the subscripts h and k denote sources (regions from which the goods are imported). Let p_{i_h} denote the unit producer (exporter) price of good i exported from country h. Denoting p the vector (set) of prices and u the utility level of the consumer, the PIGLOG (Price-Independent Generalized Logarithmic) cost function reads

$$\log C(u, p) = (1 - u) \log A(p) + u \log B(p), \tag{8.1}$$

where $A(p)$ and $B(p)$ are parametric expressions of prices:

$$\begin{aligned} \log A(p) = \alpha_0 &+ \sum_i \sum_h \log p_{i_h} \\ &+ 1/2 \sum_i \sum_j \sum_h \sum_k \gamma^*_{i_h j_k} \log p_{i_h} \log p_{j_k}, \end{aligned} \tag{8.2}$$

$$\log B(p) = \log A(p) + \beta_0 \prod_i \prod_h p_{i_h}^{\beta_{i_h}}. \tag{8.3}$$

Differentiating the (logarithmic) cost function above with respect to p and u and rearranging, it can be shown that the share of demand for good i from source h, denoted w_{i_h}, is

$$w_{i_h} = \alpha_{i_h} + \sum_j \sum_k \gamma_{i_h j_k} \log p_{j_k} + \beta_{i_h} \log(E/P^*), \tag{8.4}$$

where

$$\log P^* = \alpha_0 + \sum_i \sum_h \log p_{i_h} + 1/2 \sum_i \sum_j \sum_h \sum_k \gamma^*_{i_h j_k} \log p_{i_h} \log p_{j_k}. \tag{8.5}$$

The price index P^* is typically involving all prices from all possible sources, and it can be replaced by a weighted price index using as weights the shares in total demand of the goods, and denoted P^T in the following. The share equations defined above are valid under very general circumstances for a wide variety of goods. Parameter $\gamma_{i_h j_k}$ for instance captures the sensitivity of the import share of good i from source h with respect to the price of any good j from any source k. Of course, when $i = j$ and $h = k$, we can measure the reaction of relative demand for a given good to its own price.

If we were to use this equation directly, the number of parameters would be tremendous, as it increases both with the number of goods and the number of import sources. A way around this problem is to impose parametric restrictions to the model, by excluding the influence of the price of some goods (or sources) on some other goods (or sources). A first restriction concerns source differentiation.

RESTRICTION 8.1. Cross-price effects are not source differentiated between products, but are source differentiated within a product:

$$\gamma_{i_h j_k} = \gamma_{i_h j} \quad \forall k \in j, j \neq i. \tag{8.6}$$

For example, the country import demand for European dairy products may have a source-differentiated cross-price effect for dairy products from other sources (North America, Rest of the World), but cross-price responses to non-dairy products are not source-differentiated. This restriction implies an absence of substitutability between products of different nature and different origin.

With this restriction, the model is denoted RSDAIDS (Restricted Source Differentiated AIDS) and reads:

$$w_{i_h} = \alpha_{i_h} + \sum_k \gamma_{i_h k} \log p_{i_k} + \sum_{j \neq i} \gamma_{i_h j} \log p_j + \beta_{i_h} \log\bigl(E/P^T\bigr), \tag{8.7}$$

where $\gamma_{i_h k}$ denotes the price coefficients of good i from different source h, and p_j is a price index for all goods other than i, defined as a Tornqvist share-weighted price index.

With Restriction 8.1, the share of good i imported from source h is seen to depend on prices of the same good from all sources (coefficients

$\gamma_{i_h k}$), but also on price indexes for all other goods. This may create estimation problems if the number of goods is prohibitive. Moreover, there are reasons to believe the substitutability patterns between goods may be limited in many cases (for instance, food and machinery, precious stones and chemical products, etc.). For this reason, we further restrict the model to be estimated for import shares.

RESTRICTION 8.2. Cross-price effects are source differentiated within a product, but are not differentiated across different products. We have the final specification for the share equations:

$$w_{i_h} = \alpha_{i_h} + \sum_k \gamma_{i_h k} \log p_{i_k} + \beta_{i_h} \log\bigl(E/P^T\bigr). \tag{8.8}$$

A last set of restrictions does not actually depend on the ones imposed above on differentiation patterns, but is typically imposed in order to be consistent with the definition of the expenditure function C defined above. First, coefficients for cross-price effects should be symmetric across equations. For instance, the marginal impact of a change in the price of good i from source h on the share of good i imported from source k should be the same as the marginal impact of a change in the price of good i from source k on the share of good i imported from source h.

RESTRICTION 8.3 (Symmetry).

$$\gamma_{i_h k} = \gamma_{i_k h} \quad \forall i, h, k. \tag{8.9}$$

Second, we impose homogeneity in price for the expenditure function C, as follows:

RESTRICTION 8.4 (Homogeneity in prices).

$$\sum_h \alpha_{i_h} = 1 \quad \forall i, \qquad \sum_k \gamma_{i_h k} = 0 \quad \forall i, h. \tag{8.10}$$

Once the parametric share equations are estimated, it is possible to infer many substitutability patterns and price effects (see Green and Alston, 1990 for a survey on elasticities in the AIDS model). The Marshallian (uncompensated) own-price elasticity, measuring the change in the quantity demanded for good i from source h resulting from a change in its own price, is:

DEFINITION 8.1 (Own-price elasticity of demand).

$$\varepsilon_{i_h i_h} = -1 + \gamma_{i_h h}/w_{i_h} - \beta_{i_h}. \tag{8.11}$$

The second one is the Marshallian (uncompensated) cross-price elasticity, measuring the change in quantity demanded for good i from source h resulting from a change in the price of the same good but from a different source, k:

DEFINITION 8.2 (Cross-price elasticity of demand).

$$\varepsilon_{i_h i_k} = \gamma_{i_h k}/w_{i_h} - \beta_{i_h}(w_{i_k}/w_{i_h}). \tag{8.12}$$

The third and last one concerns the expenditure elasticity, i.e., the percent change in total demand for good i when total expenditure on all goods changes:

DEFINITION 8.3 (Expenditure elasticity).

$$\eta_{i_h} = 1 + \beta_{i_h}/w_{i_h}. \tag{8.13}$$

The import own-price and cross-price elasticities are the central objects of our empirical analysis. With the former we can predict the change in the quantity demanded for any given commodity from any source when its price increases or decreases. With the latter it is possible to assess the degree of trade diversion patterns associated with each good. For instance, an increase in the price of imported European goods relative to North American ones is likely to lead to an increase in imports from North America and a decrease in imports from Europe, but it may also have an impact on goods imported from countries or from the rest of the world. This is because, among other things, additional disposable revenue is obtained following this decrease in the price of imported goods from Europe, and imports from the rest of the world of competing goods that were not considered for importing may also increase.

The use of the RSAIDS model for modeling import demand can in principle be performed on a time series of observations for various commodities from different regions, or even using a cross section of products. However, to limit the number of parameters even further, it is often necessary to impose an additional separability restriction and consider a subgroup of products only. The approach used in this paper is to concentrate on a limited number of product categories, namely agricultural products, and to assume that expenditure decisions on these commodities can be separated from other import decisions regarding non-agricultural goods. We further assume separability between agricultural product groups. In other words, we consider that the expenditure variable defined above is the sum of import values from all possible sources, for the agricultural category under consideration. Therefore, when interpreting expenditure or income effect within the AIDS context, this point should be remembered, as expenditure in our sense may not vary in line with total expenditure. Two points can be advanced to justify such a restriction. First, we are mostly interested in competition among export regions and countries rather than substitution among products *and* competition. Second, as will be discussed below, our data are very disaggregate and are available at the individual importer level. Since importers in agricultural commodities are expected to be rather specialized, this limiting aspect of the separability assumption may be reduced.

8.3. The multi-way unbalanced error-component panel data model

Consider the following import share panel data equation:

$$w_{iht} = X_{iht}\beta + u_{iht},$$
$$i = 1, \ldots, L; h = 1, \ldots, H;\; t = 1, \ldots, T, \tag{8.14}$$

where X is the matrix of agricultural prices and expenditure (in log), and β the slope parameter vector to be estimated. i, h and t are the commodity, country and period index respectively. L is the number of goods ever imported from any given country, and H is the total number of countries. The total number of observations is N and the sample is unbalanced, i.e., the number of available observations (time periods) for good i and a given country h is not necessarily constant. In fact, it can even be zero if a country does not export a given commodity at all.

However, the multi-way error components model defined above is not nested in a strict sense, because H (resp. L) is not indexed by i (resp. h). For the import share model, a nested structure would imply that exporters are completely exclusive in the determination of their products, and that countries are fully specialized in exports. This is of course not true in practice, as it would rule out competition on export markets for homogeneous products.

We consider the following expression for the error term:

$$u_{iht} = \alpha_i + \gamma_h + \lambda_t + \varepsilon_{iht},$$
$$i = 1, \ldots, L; h = 1, \ldots, H;\; t = 1, \ldots, T, \tag{8.15}$$

where α_i, γ_h and λ_t are i.i.d. error components. Random effects α_i and γ_h are included for capturing unobserved heterogeneity in demand shares, not explained by prices, and related to commodity and country respectively, whereas λ_t is picking up time-effects. The error term ε is i.i.d. across all dimensions (product, country, period), with variance denoted σ_ε^2.

In this framework, the commodity effect α_i is capturing unobserved attributes of commodity i, irrespective of the country and constant across time periods, and which are systematic in the demand share. Likewise, γ_h is the country-specific heterogeneous intercept capturing the idiosyncratic component of imports from country h, independent from commodities and time periods. It is tempting to interpret such unobserved terms as quality measures that may be independent from prices, but it has to be remembered that the AIDS demand system places a particular restriction on the intercept term in the share equation. Hence, in our case, unobserved quality would enter the model through heterogeneous own-price slopes in the expenditure function, a specification which closely resembles the one proposed by Crozet and Erkel-Rousse (2004).

In matrix form, the three-way ECM (Error Component Model) can be written

$$Y = X\beta + u, \quad \text{with } u = \Delta_1\alpha + \Delta_2\gamma + \Delta_3\lambda + \varepsilon, \tag{8.16}$$

where $\alpha = (\alpha_1, \ldots, \alpha_L)'$, $\gamma = (\gamma_1, \ldots, \gamma_H)'$ and $\lambda = (\lambda_1, \ldots, \lambda_T)'$.

Matrices Δ_k, $k = 1, 2, 3$, contain dummy variables equal to 1 if observation (i, h, t) is relevant for the group.

As mentioned above, it is possible that no observation are present for a given combination (i, h). Davis (2002) points out that the original model does not need be additive multi-way or nested, but can consist in a more complex combination of the two usual specifications. The only requirement is that matrices of dummy variables have full column-rank.

8.3.1. The fixed effects model

Consider the fixed effects estimator first. If we suspect random effects to be correlated with components of matrix X, we can obtain a consistent estimator by wiping out these effects α, γ and λ (within-group transformation) or, equivalently, directly incorporating them in the right-hand side of the equation (LSDV, Least-Squares Dummy Variables procedure). This kind of conditional inference can be performed more easily with the within approach if the dataset is very large, and the number of dummy variables is important (when it is increasing with N for instance).

Assume N goes to infinity and denote $\Delta = (\Delta_1, \Delta_2, \Delta_3)$ the matrix of dummy variables (indicator variable matrix) associated with the three-way ECM. Matrices Δ_1, Δ_2 and Δ_3 have dimensions $N \times L$, $N \times H$ and $N \times T$ respectively. Constructing the within matrix operator with unbalanced panel in this case can be done in several stages. Let $P_A = A(A'A)^+A'$ and $Q_A = I - P_A$, where $^+$ denotes a generalized inverse, and Q_A is the projection onto the null space of matrix A. The matrix Q_A is idempotent and symmetric. Using results in Wansbeek and Kapteyn (1989) and Davis (2002), it can be shown that the required transformation obtained as:

$$\begin{aligned}
&Q_\Delta = Q_A - P_B - P_C, \\
&P_A = I - \Delta_3(\Delta_3'\Delta_3)^+\Delta_3', \qquad Q_A = I - P_A, \\
&P_B = Q_A\Delta_2(\Delta_2'Q_A\Delta_2)^+\Delta_2'Q_A, \qquad Q_B = I - P_B, \\
&P_C = Q_AQ_B\Delta_1\bigl[\Delta_1'(Q_AQ_B)\Delta_1\bigr]^+\Delta_1'Q_AQ_B, \qquad Q_C = I - P_C.
\end{aligned} \tag{8.17}$$

The within estimator is therefore defined as

$$\hat{\beta} = (X' Q_\Delta X)^{-1} X' Q_\Delta Y, \tag{8.18}$$

under the exogeneity assumption $E[X' Q_\Delta \varepsilon] = 0$.

8.3.2. The random effects model

Consider now the random effects model, with the same multi-way error-components structure defined above. We assume that α_i, γ_h and λ_t have zero mean and variance σ_1^2, σ_2^2 and σ_3^2 respectively. The full $N \times N$ variance–covariance matrix of the model is

$$\Omega = E(uu') = \sigma_\varepsilon^2 I + \sigma_1^2 \Delta_1 \Delta_1' + \sigma_2^2 \Delta_2 \Delta_2' + \sigma_3^2 \Delta_3 \Delta_3'. \tag{8.19}$$

Under the normality assumption, the log-likelihood function is

$$\log L = -\frac{N}{2} \log(2\pi) + \frac{1}{2} \log\left|\Omega^{-1}\right| - \frac{1}{2} u' \Omega^{-1} u, \tag{8.20}$$

which is maximized jointly with respect to slope parameters β and error component variances σ_1^2, σ_2^2, σ_3^2 and σ_ε^2.

Although conforming with non-linear optimization gradient-based routines, maximizing the log-likelihood may reveal cumbersome with large datasets, as Ω, an $N \times N$ matrix, needs to be inverted. In the balanced panel data case, inverting this variance–covariance matrix is made easy by the fact that the spectral decomposition of Ω involves identical block-diagonal terms associated with scalar variance components. A Fuller and Battese (1973) spectral transformation can in principle be performed directly on the dependent and the RHS variables. In the unbalanced but nested model, Baltagi *et al.* (2001) show that such a transformation is straightforward. We present below the corresponding spectral decomposition for the unbalanced, non-nested three-way error components model.

In practice, because the information matrix is block-diagonal between first- and second-moment model parameters, estimation of the random effects model is often performed in two stages to avoid optimization routines. The Feasible GLS estimator, based on initial and consistent estimates of variance components, is consistent and asymptotically equivalent to the Maximum Likelihood Estimator (MLE).

There are several methods in the literature for estimating variance components, but asymptotic properties of these estimators differ, depending on the nature of the panel dataset (balanced or unbalanced). With a balanced sample, ANOVA estimators are Best Quadratic Unbiased (BQU), and are minimum variance unbiased when disturbances are normally distributed, see Searle (1987). In the unbalanced panel case however, these ANOVA variance estimation procedures only yield unbiased estimates for error components. Baltagi *et al.* (2001) presents several extensions of ANOVA estimators originally designed for the balanced data case, to the unbalanced case with multi-way error components (Baltagi and Chang, 1994 consider an extension of this approach to the one-way unbalanced panel data model only).

In this paper, we consider two different methods for estimating variance components. The first one is the Quadratic Unbiased Estimator (QUE) analogue to the two-way Wansbeek and Kapteyn (1989) estimator, based on Within residuals obtained using Equation (8.18). Let $S_N = e'Q_{\varDelta}e, S_i = e'P_{\varDelta_i}e$ for $i = 1, 2, 3$, where e is the N-vector of Within residuals. N_1, N_2 and N_3 are the column dimensions of $\varDelta_1$, $\varDelta_2$ and $\varDelta_3$ respectively. Quadratic unbiased estimators of variance components obtain by solving the following system of equations:

$$\begin{aligned}
&E(S_N) = (N - t_1 - t_2 - t_3 + k_n)\sigma_\varepsilon^2,\\
&E(S_1) = (N_1 + k_1)\sigma_\varepsilon^2 + n\sigma_1^2 + k_{12}\sigma_2^2 + k_{13}\sigma_3^2,\\
&E(S_2) = (N_2 + k_2)\sigma_\varepsilon^2 + k_{21}\sigma_1^2 + n\sigma_2^2 + k_{23}\sigma_3^2,\\
&E(S_3) = (N_3 + k_3)\sigma_\varepsilon^2 + k_{31}\sigma_1^2 + k_{32}\sigma_2^2 + n\sigma_3^2,
\end{aligned}$$

where $k_N = \operatorname{rank}(X), k_i = \operatorname{tr}[(X'Q_{\varDelta}X)^{-1}X'P_{\varDelta_i}X], k_{ij} = \operatorname{tr}[\varDelta_j' P_{\varDelta_i}\varDelta_j]$, $i, j = 1, 2, 3, t_1 = \operatorname{rank}(A), t_2 = \operatorname{rank}(B), t_3 = \operatorname{rank}(C)$, and $A = \varDelta_3, B = Q_A\varDelta_2, C = Q_AQ_B\varDelta_1$.

The second variance estimation procedure we consider is the Minimum Norm Quadratic Unbiased Estimator (MINQUE) proposed by Rao (1971) for a very general error component model. Letting $\theta = (\sigma_\varepsilon^2, \sigma_1^2, \sigma_2^2, \sigma_3^2)$, the MINQUE estimator of variance components, conditional on parameter estimates $\hat{\beta}$ is defined by $\hat{\theta} = S^{-1}u$, where

$$\begin{aligned}
&S = \{\operatorname{tr}(V_iRV_jR)\}_{i,j}, \quad i, j = 0, 1, 2, 3,\\
&u = \{y'RV_iRy\}_i, \quad i = 0, 1, 2, 3,\\
&V_i = \varDelta_i\varDelta_i', \quad i = 0, 1, 2, 3,\ \varDelta_0 = I,\\
&R = \varOmega^{-1}\big[I - X\big(X'\varOmega^{-1}X\big)^{-1}X'\varOmega^{-1}\big].
\end{aligned}$$

QUE estimates presented above can be used as initial values for computing R. As is well known in applied work involving random effects specifications, variance estimates can in some instances be negative, even in simpler cases than ours. There does not seem to be fully satisfactory ways to overcome this problem in practice, apart from considering a restricted error component specification (e.g., considering a two-way instead of a three-way model).

Once variance components are estimated, the inverse of the variance–covariance Ω can be constructed to compute the Feasible GLS estimator $\hat{\beta} = (X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}Y$. Davis (2002) presents a convenient way of constructing this inverse matrix without inverting any $n \times n$ term. This is based on the fact that the inverse of Ω can be written as $\Omega^{-1} = \sigma_{\varepsilon}^{-2}(I_N + DD')^{-1}$, where

$$\begin{aligned}
&(I_N + DD')^{-1} = \widetilde{Q}_{\Delta_2,\Delta_3} - \widetilde{Q}_{\Delta_2,\Delta_3}\Delta_1 W_1^{-1}\Delta_1'\widetilde{Q}_{\Delta_2,\Delta_3},\\
&\widetilde{Q}_{\Delta_2,\Delta_3} = \widetilde{Q}_{\Delta_3} - \widetilde{Q}_{\Delta_3}\Delta_2 W_2^{-1}\Delta_2'\widetilde{Q}_{\Delta_3},\\
&\widetilde{Q}_{\Delta_3} = I_N - \Delta_3(I + \Delta_3'\Delta_3)^{-1}\Delta_3',\\
&W_1 = I + \Delta_1'\widetilde{Q}_{\Delta_2,\Delta_3}\Delta_1, \qquad W_2 = I + \Delta_2'\widetilde{Q}_{\Delta_3}\Delta_2.
\end{aligned}$$

Hence, only matrices with rank $\text{cols}(\Delta_1) = N_1, \text{cols}(\Delta_2) = N_2$ and $\text{cols}(\Delta_3) = N_3$ need to be inverted.

Interestingly, the inverse of the variance–covariance matrix can also be obtained using a spectral decomposition (such a decomposition was obtained by Baltagi *et al.*, 2001 in the unbalanced nested case). Compared to their case however, because our model is not nested, the redefinition of matrices Z_2 and Z_3 and a change in the notation of the effect dimensions are needed, to be able to use the technique (and notation) developed in Baltagi *et al.* (2001).

Consider again the three-way error components model (8.15), where now H and T are indexed by i:

$$\begin{aligned}
&u_{iht} = \alpha_i + \gamma_h + \lambda_t + \varepsilon_{iht},\\
&\quad i = 1, \ldots, L;\ h = 1, \ldots, H_i;\ t = 1, \ldots, T_i,
\end{aligned}$$

and define the following vectors:

$$\begin{aligned}
&\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_L)',\\
&\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{H_1}, \ldots, \gamma_1, \gamma_2, \ldots, \gamma_{H_L})',\\
&\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{T_1}, \ldots, \lambda_1, \lambda_2, \ldots, \lambda_{T_L})',\\
&\varepsilon = (\varepsilon_{111}, \varepsilon_{112}, \ldots, \varepsilon_{11T_1}, \ldots, \varepsilon_{LH_L1}, \varepsilon_{LH_L2}, \ldots, \varepsilon_{LH_LT_L})'.
\end{aligned}$$

We have $N = \sum_{i=1}^{L} H_i T_i$.

Unbalancedness in the country and time dimensions is therefore taken care of by duplicating effects γ and λ according to their relevance for a given commodity level (i).

The error term in matrix form reads:

$$u = Z_1\alpha + Z_2\gamma + Z_3\lambda + \varepsilon, \tag{8.21}$$

where

$$Z_1 = \text{diag}(e_{H_i} \otimes e_{T_i}), \qquad Z_2 = \text{diag}(I_{H_i} \otimes e_{T_i}),$$

$Z_3 = \mathrm{diag}(e_{H_i} \otimes I_{T_i})$,

and e_{H_i} (resp. e_{T_i}) is a H_i (resp. T_i) vector of ones.

By $\mathrm{diag}(e_{H_i} \otimes e_{T_i})$ we mean $\mathrm{diag}(e_{H_1} \otimes e_{T_1}, e_{H_2} \otimes e_{T_2}, \ldots, e_{H_L} \otimes e_{T_L})$, a block-diagonal matrix, and accordingly for terms Z_2 and Z_3.

The variance–covariance matrix is now

$$\begin{aligned} \Omega &= \sigma_1^2 Z_1 Z_1' + \sigma_2^2 Z_2 Z_2' + \sigma_3^2 Z_3 Z_3' + \sigma_\varepsilon^2 \,\mathrm{diag}(I_{H_i} \otimes I_{T_i}) \\ &= \mathrm{diag}\big[\sigma_1^2 (M_{H_i} \otimes M_{T_i}) + \sigma_2^2 (I_{H_i} \otimes M_{T_i}) + \sigma_3^2 (M_{H_i} \otimes I_{T_i}) \\ &\quad + \sigma_\varepsilon^2 (I_{H_i} \otimes I_{T_i})\big] = \mathrm{diag}(\Lambda_i), \end{aligned} \tag{8.22}$$

where, following Baltagi *et al.* (2001) notation, $M_{H_i} = e_{H_i} e_{H_i}'$ and $M_{T_i} = e_{T_i} e_{T_i}'$.

We have

$$\begin{aligned} \Lambda_i &= \sigma_1^2 H_i T_i (\overline{M}_{H_i} \otimes \overline{M}_{T_i}) + \sigma_2^2 T_i (I_{H_i} \otimes \overline{M}_{T_i}) \\ &\quad + \sigma_3^2 H_i (\overline{M}_{H_i} \otimes I_{T_i}) + \sigma_\varepsilon^2 (I_{H_i} \otimes I_{T_i}), \end{aligned} \tag{8.23}$$

where $\overline{M}_{H_i} = M_{H_i}/H_i$ and $\overline{M}_{T_i} = M_{T_i}/T_i$.

Letting $E_{H_i} = I_{H_i} - \overline{M}_{H_i}$ and $E_{T_i} = I_{T_i} - \overline{M}_{T_i}$, the spectral decomposition of Λ_i can be written as

$$\Lambda_i = \pi_{1i} Q_{1i} + \pi_{2i} Q_{2i} + \pi_{3i} Q_{3i} + \pi_{4i} Q_{4i}, \tag{8.24}$$

where

$$\begin{aligned} &\pi_{1i} = \sigma_\varepsilon^2, \qquad \pi_{2i} = T_i \sigma_2^2 + \sigma_\varepsilon^2, \qquad \pi_{3i} = H_i \sigma_3^2 + \sigma_\varepsilon^2, \\ &\pi_{4i} = H_i T_i \sigma_1^2 + T_i \sigma_2^2 + H_i \sigma_3^2 + \sigma_\varepsilon^2, \\ &Q_{1i} = (E_{H_i} \otimes I_{T_i})(I_{H_i} \otimes E_{T_i}), \qquad Q_{2i} = (E_{H_i} \otimes \overline{M}_{T_i}), \\ &Q_{3i} = (\overline{M}_{H_i} \otimes E_{T_i}), \qquad Q_{4i} = (\overline{M}_{H_i} \otimes \overline{M}_{T_i}). \end{aligned}$$

It is easy to show in particular that $Q_1 Z_1 = Q_1 Z_2 = Q_1 Z_3 = 0$, where$Q_1 = \mathrm{diag}(Q_{1i})$, hence satisfying the orthogonality conditions for the spectral decomposition to apply. Based on consistent (not necessarily unbiased) estimates of the variance components, a Feasible GLS estimation through data transformation using scalar expressions only can be performed.

Note finally that in the presence of serial correlation in the ε's, using QUE or MINQUE variance components estimators may be problematic. In this case, $\sigma_\varepsilon^2 I$ would become $\sigma_\varepsilon^2 (I \otimes \Sigma_\varepsilon)$, with Σ_ε the serial correlation matrix whose parameter(s) need to be estimated. Further research should be devoted to extend the unbalanced three-way error component model to allow for serial correlation in the disturbances.

8.3.3. Specification tests[2]

With multi-way error components models, the issue of specification tests is even more important for empirical purposes, as far as the presence of effects is concerned. Misspecification in the structure of the error term will have two kinds of consequences for parametric estimation. First, under a random effects model with exogeneity of explanatory variables assumed, standard-error estimates will be biased, although point estimates will remain consistent. Second, in a fixed effects context, the consequence of such a misspecification will be worse if omitted unobserved heterogeneity terms are correlated with explanatory variables.

In our case, assuming for example a one-way model with commodity effects only (α_i in the notation above) whereas the true model entails a country effect (γ_h) as well, will produce inconsistent slope estimates in import demand share equations if there exists a systematic country-wise component correlated with import prices. Not only GLS but also Within estimates are expected to be biased in this case, and an Hausman-type exogeneity test between the two sets of estimates under the misspecified one-way model will not provide indication of such bias.

Two types of tests need to be performed on the model. First, it is necessary to check for the validity of assumptions made on the structure of the multi-way error components model. A natural option is to use the Lagrange Multiplier test statistic based on components of the log-likelihood evaluated at parameter estimates. This option is motivated by the equivalence between GLS and Maximum Likelihood estimation, under the exogeneity assumption. Hence, it is obvious that the validity of the test statistic will depend crucially on this assumption under the random effects specification.

Formally, the Lagrange Multiplier is written as $\widetilde{D}'\widetilde{F}^{-1}\widetilde{D}$ where $\widetilde{D}$ and $\widetilde{F}$ respectively denote the restricted score vector and information matrix. When testing for the significance of random effects, i.e., testing for the nullity of their associated variances, we let $\theta = (\sigma_\varepsilon^2, \sigma_1^2, \sigma_2^2, \sigma_3^2)$ and consider $\widetilde{D} = D(\tilde{\theta})$, $\widetilde{F} = F(\tilde{\theta})$ where $\tilde{\theta}$ is the constrained vector of variances and

$$\begin{aligned} D(\theta) &= \left\{ \frac{\partial \log L}{\partial \theta_r} \right\}_r \\ &= \left\{ -\frac{1}{2} \operatorname{tr}\left[\Omega^{-1} \left(\frac{\partial \Omega}{\partial \theta_r} \right) \right] + \frac{1}{2} \left[u' \Omega^{-1} \left(\frac{\partial \Omega}{\partial \theta_r} \right) \right] \Omega^{-1} u \right\}_r, \end{aligned}$$

[2] This section relies on the discussion (and notation) in Davis (2002) about variance components significance tests.

$$F(\theta) = E\left\{-\frac{\partial^2 \log L}{\partial\theta_r \partial\theta_s'}\right\}_{(r,s)} = \left\{\frac{1}{2}\operatorname{tr}\left[\varOmega^{-1}\left(\frac{\partial\varOmega}{\partial\theta_r}\right)\varOmega^{-1}\left(\frac{\partial\varOmega}{\partial\theta_s}\right)\right]\right\}_{(r,s)}.$$

If one wishes for example to test for the absence of all effects, we set $\tilde{\theta} = (\sigma_\varepsilon^2, 0, 0, 0)$ and consider the restricted variance–covariance matrix $\widetilde{\varOmega} = \tilde{\sigma}_\varepsilon^2 I$, where $\tilde{\sigma}_\varepsilon^2 = \frac{1}{N}\tilde{u}'\tilde{u}$ is the restricted variance from OLS residuals. Letting $t_1 = \operatorname{tr}(\varDelta_1\varDelta_1')$, $t_2 = \operatorname{tr}(\varDelta_2\varDelta_2')$, $t_3 = \operatorname{tr}(\varDelta_3\varDelta_3')$, the LM statistic under H_0: $\theta = \tilde{\theta}$ is given by

$$\mathrm{LM} = \left(-\frac{N}{2\tilde{\sigma}_\varepsilon^2}\right)^2 \left(2\tilde{\sigma}_\varepsilon^4\right)\tilde{\tilde{D}}'\tilde{\tilde{F}}^{-1}\tilde{\tilde{D}} = \frac{N^2}{2}\tilde{\tilde{D}}'\tilde{\tilde{F}}^{-1}\tilde{\tilde{D}}, \tag{8.25}$$

where

$$\tilde{\tilde{D}} = \begin{pmatrix} 0 \\ \frac{t_1}{N} - \frac{\tilde{u}'(\varDelta_1\varDelta_1')\tilde{u}}{\tilde{u}'\tilde{u}} \\ \frac{t_2}{N} - \frac{\tilde{u}'(\varDelta_2\varDelta_2')\tilde{u}}{\tilde{u}'\tilde{u}} \\ \frac{t_3}{N} - \frac{\tilde{u}'(\varDelta_3\varDelta_3')\tilde{u}}{\tilde{u}'\tilde{u}} \end{pmatrix},$$

$$\tilde{\tilde{F}} = \begin{bmatrix} N & \operatorname{tr}(\varDelta_1\varDelta_1') & \operatorname{tr}(\varDelta_2\varDelta_2') & \operatorname{tr}(\varDelta_3\varDelta_3') \\ & \operatorname{tr}(\varDelta_1\varDelta_1'\varDelta_1\varDelta') & \operatorname{tr}(\varDelta_1\varDelta_1'\varDelta_2\varDelta_2') & \operatorname{tr}(\varDelta_1\varDelta_1'\varDelta_3\varDelta_3') \\ & & \operatorname{tr}(\varDelta_2\varDelta_2'\varDelta_2\varDelta_2') & \operatorname{tr}(\varDelta_2\varDelta_2'\varDelta_3\varDelta_3') \\ & & & \operatorname{tr}(\varDelta_3\varDelta_3'\varDelta_3\varDelta_3') \end{bmatrix}.$$

Finally, to test for the validity of our error-component specification, we can easily compute the LM test statistic under various nullity assumptions on individual variances. In other terms, we may test for the joint significance of a subset of variances only, or test for the variance of a particular effect to be 0, or finally test that the 3 variance components are 0. This is particularly important when deciding whether a one-way of a two-way model should be preferred, that is, in our case, if country-specific effects should be accounted for in addition to product-specific individual effects.

The second type of specification checks concerns Hausman-type exogeneity tests, where GLS are compared to fixed effects estimates. Under the assumption that the first specification analysis above has properly identified the genuine error components structure, we test for the lack of correlation between effects α_i, γ_h and λ_t on the one hand, and explanatory variables on the other, by considering the null hypothesis H_0: $E(X'\varOmega^{-1}u) = 0$. This condition can be tested by verifying whether components in the original model disturbance that are orthogonal to $Q_\varDelta$, are correlated with X or not, in exactly the same way as in the one-way model, when additional orthogonality conditions are imposed on the Between component.

Table 8.1. Weighted average import tariff rates

Chapter		Average customs tariff rate (percent)
2	Meat and edible meat offal	5.01
4	Dairy products; birds' eggs; natural honey	13.68
10	Cereals, products of the milling industry	1.31
15	Animal or vegetable fats and oils	12.92
17	Sugars and sugar confectionery	9.45

Source: Lebanese Ministry of Finance – Customs Administration.

8.4. The data

Data were made available from the Lebanese Customs administration for the years 1997–2002. Original data records consist of exhaustive daily transactions (both exports and imports) and contain information on: day, month and year of the transaction; country of origin (imports) or destination (exports); preferential or trade agreement tariffs; net quantity of commodity imported/exported (either net weight or number of units); amount of the transaction (in Lebanese pounds and USD).

We only selected the main agricultural import categories from the database, corresponding to chapters 2 (meat), 4 (milk & dairy products), 10 (cereals), 15 (animal & vegetable fats & oils) and 17 (sugar) in the International Harmonized System (HS) classification. For each transaction, we computed unit prices by dividing the import transaction amount by the number of units when applicable, or by net weight. These unit prices before application of customs tariffs (but including cost, insurance and freight, CIF) were then converted to unit prices inclusive of customs duties, using average weighted tariff rates. Table 8.1 reports values for these tariff rates, based on the 1999 import values. Except for cereals (chapter 10), imported agricultural goods are associated with rather high customs duty rates, but these rates are more or less similar to other categories: the average weighted rate for non-agricultural imports to Lebanon was 12.20 percent in 1999 and fell to 6.23 percent in 2001.

European imports for the products described above were selected from the database, and the monthly average import price was computed by product level (level HS8 in the harmonized system) for each European country. The choice of the HS8 level as the base unit for the empirical analysis is motivated by the need to preserve a reasonable level of homogeneity for commodities.

Import price is then associated with the import share corresponding to the European country and the commodity group, which is computed for the corresponding time period (month) by summing all imports for this product from the same European country, and dividing by the sum of total imports for this product category (chapter), from all regions of the world (E_t):

$$w_{i_h t} = p_{i_h t} Q_{i_h t} \Big/ \sum_i \sum_h p_{i_h t} Q_{i_h t} = p_{i_h t} Q_{i_h t} / E_t. \tag{8.26}$$

Hence, for every chapter of products (2, 4, 10, 15 or 17), the sum of shares will be 1 for any given product and time period, where a single share can be either from one of the 15 European countries, or from one of the 3 other regions we have defined. These regions are denoted AR (Arab and Regional countries), AM (North and South America) and ROW (Rest of the World).

Monthly price indexes are computed for all agricultural products under consideration, for these three export regions. These regions are defined in a narrow economic sense in the case of the European Union, and in a more geographic sense for AR (Arab and Regional countries) and AM (North and Latin America).

The country classification is the following:

- EU (European Union): Austria, Belgium, Germany, Denmark, Spain, Finland, France (including Andorra, Guadeloupe, Martinique and Réunion), Great-Britain (including Gibraltar), Greece, Ireland, Italy, Luxemburg, The Netherlands, Portugal, Sweden;
- AR (Arab and Regional countries): Algeria, Morocco, Tunisia, Libya, Iraq, Jordan, Kuwait, Arab Emirates, Bahrain, Brunei, Egypt, Iran, Oman, Qatar, Saudi Arabia, Sudan, Syria, Turkey;
- AM (North and Latin America): United States of America, Canada, Argentina, Bolivia, Brazil, Bahamas, Chile, Colombia, Costa Rica, Dominican, Ecuador, Guatemala, Honduras, Mexico, Panama, Peru, Paraguay, Trinidad and Tobago, Uruguay, Venezuela;
- ROW (Rest of the World): all other countries excluding Lebanon.

For each good i (defined at the HS8 level) and month t, we construct an import-share-weighted Tornqvist price index as follows:

$$\log p_{it} = \sum_k \left(w_{i_k t} + w_k^0\right) \log\left(p_{i_k t} / p_k^0\right), \tag{8.27}$$

where $w_{i_k t}$ denotes the share of good i imported from source k (AR, AM or ROW) at time t, $p_{i_k t}$ is the price of good i imported from source k; w_k^0 and p_k^0 are the average (over all time periods) import share and price for source k, respectively. The Tornqvist price index is the approximation of the Divisia index, and it is chosen because it is consistent with the AIDS demand system.

To control for strategic behavior from European competitors, we also compute the average import price by product, for all European imports excluding the country under consideration. The price index in this case is also of the Tornqvist form presented above, and will be denoted P_{EU} in the following. Depending on the sign of the associated coefficient of this cross-price effect, substitutability or complementarity patterns between European countries can be identified. This aspect of the model is of course related to the availability of highly disaggregated data, as Lebanese imports from different countries and different products are observed on a high frequency basis.

The need for competitor prices at each commodity level and for each time period (month) implies that a significant amount of data is lost because of missing variables. This problem, however, is mainly due to the lack of observations for some time periods on regions AR and AM exports to Lebanon. When this is the case, the time period is entirely lost. However, when no import data are available for some European countries, the estimation procedure accommodates for this, because the model is precisely unbalanced in this regard: European import shares depend on countries and commodities, for which the number of available (time) observations differ. In total, there are 3133 monthly observations, for 15 countries (in the European Union), 51 products (HS8 level) and 72 months.

The share equations corresponding to the Restricted Source Differentiated AIDS demand system imply not only price indexes for various import sources as explanatory variables, but also the logarithm of expenditure over the overall price index, $\log(E_t/P^T)$. This expenditure is an endogenous variable in the statistical and economic sense, as it depends on the whole price system including region-specific import unit price indexes. For this reason, it is common practice in applied demand analysis to replace discounted expenditure by a prediction computed from instruments such as time dummies and a selection of prices (see Andayani and Tilley, 1997), in order to resolve this endogeneity problem. We estimate an autoregressive process for this expenditure *in the considered agricultural product category* (*chapter*) *only* (in log), incorporating yearly dummies (from 1998 to 2002) as well, and retain its estimated value, denoted EXP_{it}. This linear prediction is then used in place of the original variable in the share equations. With such a procedure, parameter estimates are expected to be consistent. As mentioned above, we restrict the demand model in such a way that no substitution patterns are allowed between agricultural goods and non-agricultural commodities, as well as between different agricultural categories, to keep the number of parameters to a minimum.

Average regional import shares over the period 1997–2002 are reported in Table 8.2, for each of the 6 commodity chapters. One can see from

Table 8.2. Average import shares, by region – 1997 to 2002 average

Chapter (HS2)		Europe	Arab & Regional	America	ROW
2	(Meat)	0.2484	0.0731	0.2630	0.4153
4	(Dairy products)	0.4062	0.1436	0.0783	0.3717
10	(Cereals)	0.2553	0.1827	0.1734	0.3885
15	(Animal & veg. fat)	0.2630	0.3329	0.1716	0.2323
17	(Sugar)	0.3914	0.2582	0.1170	0.2332

this table that European imports are the most important in relative terms for chapter 4 (milk & dairy products) only. Imports from the rest of the world remain more significant especially for meat (chapter 2) and cereals (chapter 10). Interestingly, Arab and Regional countries have the highest average import share for fats and oils. Given the importance of these commodities particularly for home cooking in Southern Europe and the Mediterranean Sea, these countries are expected to compete more on these products with the European Union. American imports do not appear to possess a dominant situation in either commodity groups. For meat however, their import share is higher than the European ones on average.

To assess the degree of unbalancedness in the data, we compute the following measure for each component:

$$r_i = \frac{N_i^2}{\mathrm{tr}[(\Delta_i'\Delta_i)^{-1}]\mathrm{tr}(\Delta_i'\Delta_i)}, \quad i = 1, 2, 3, \tag{8.28}$$

where n_i is the column dimension of Δ_i. The expression r_i takes the value of 1 when the data are exactly balanced in all other dimensions (other than i), while a value approaching 0 indicates a severe degree of unbalancedness. In our case where $i = 1$ corresponds to product, $i = 2$ to country and $i = 3$ corresponds to time, we have

Chapter 2 – Meat	$r_1 = 0.0718$; $r_2 = 0.3899$; $r_3 = 0.9826$;
Chapter 4 – Milk and Dairy	$r_1 = 0.0840$; $r_2 = 0.2001$; $r_3 = 0.9729$;
Chapter 10 – Cereals	$r_1 = 0.1746$; $r_2 = 0.2881$; $r_3 = 0.9873$;
Chapter 15 – Fats	$r_1 = 0.0337$; $r_2 = 0.0983$; $r_3 = 0.9925$;
Chapter 17 – Sugar	$r_1 = 0.0531$; $r_2 = 0.3012$; $r_3 = 0.9872$.

The measure of unbalancedness associated with time (r_3) indicates that only a very small proportion of European countries and products are not present every year in the sample over the period 1997–2002. However, r_2 reveals that the degree of unbalancedness as far as countries are concerned is significant, with measures ranging from 0.09 to 0.39. Finally, as expected, the highest degree of unbalancedness is associated with products, because of the relative degree of specialization in agricultural products exported from the European Union. Values of r_1 by chapter seem to indicate that Cereals are exported by a more significant subset of countries (with value 0.17), while Fats are more limited in the range of exporting countries (with value 0.03).

8.5. Estimation results

The demand share equations for the 5 product categories have the following form:

$$\begin{aligned} w_{i_hkt} = {} & \beta_{0k} + \beta_{1k}\log(P_{i_hkt}/P_{\mathrm{ROW},ikt}) + \beta_{2k}\log(P_{\mathrm{EU},ikt}/P_{\mathrm{ROW},ikt}) \\ & + \beta_{3k}\log(P_{\mathrm{AR},ikt}/P_{\mathrm{ROW},ikt}) + \beta_{4k}\log(P_{\mathrm{AM},ikt}/P_{\mathrm{ROW},ikt}) \\ & + \beta_{6k}\log \mathrm{EXP}_t + u_{i_hkt}, \end{aligned} \tag{8.29}$$

where index k denotes the commodity group, $k = 2, 4, 10, 15, 17$, i denotes product, h is the European country associated with i and t is the time period. p_{i_hkt} denotes the individual import price for good i and from European country j, whereas PEU_{ikt} denotes the price index for the same good being exported from all other European competitors (i.e., from European countries k different from European country h). Observations for different commodity groups (chapters) are not pooled, so that there are 5 sets of parameters to estimate. Linear homogeneity of expenditure in agricultural goods is imposed by dividing all prices by the price index for the Rest of the World (ROW) region. The total number of observations is $N = 3133$ for the 5 different commodity groups.

The first stage in the estimation procedure is to check for the validity of the error components model, regarding the number of components in the multi-way structure (product, country, time). For this, we estimate Equation (8.13) with the random effects specification, to obtain estimates of variance components $\theta = (\sigma_\varepsilon^2, \sigma_1^2, \sigma_2^2, \sigma_3^2)$. Lagrange Multiplier test statistics are then computed for different model specifications: (product, year), (country, year), (product, country), (no effects), by using the expression $\widetilde{D}'\widetilde{F}^{-1}\widetilde{D}$ under alternative variance restrictions.

As mentioned before, this approach has the advantage of avoiding maximizing the log-likelihood with respect to variance components and slope parameters. The second step is to check for the validity of the random effects specification, which is performed by computing a Hausman test statistic for the comparison between GLS and Within (fixed effects) estimates. As is well known, if the random effects specification were to be invalidated by this exogeneity test, then GLS would not be consistent, and the LM test statistics would not be valid either. Results are presented in Tables 8.3 to 8.7 for the 5 chapters under consideration: Meat (Table 8.3, chapter 2), Milk and Dairy products (Table 8.4, chapter 4), Cereals (Table 8.5, chapter 10), Fats (Table 8.6, chapter 15) and Sugar (Table 8.7, chapter 17). For each product category (chapter), we present OLS, Within (fixed effects) and GLS parameter estimates under the full three-way model specification. For the random effects specification, two versions of GLS are computed: GLS with QUE variance components and GLS with MINQUE variance components.[3] In the case of the latter, initial values for computing variance components are the Wansbeek–Kapteyn QUE estimates. As a measure of fit associated with each estimation procedure, we also report the Root Mean Square Error (RMSE).

It can be seen first that in all cases, when using MINQUE variance estimates, the Hausman test statistic does not reject the random effects specification (although for the case of Meat, in Table 8.3, the p-value associated with the Hausman test statistic is only slightly above the 5 percent level). Hence, the test statistic is not in favor of rejecting the null hypothesis that $E[X'\Omega^{-1}u] = 0$, indicating that random effects could be considered. The value of the test statistic leads however to a different conclusion for only 1 chapter out of 5 (Fats), where GLS-QUE is rejected in favor of the Within, whereas the random effects specification is not rejected when using GLS-MINQUE.

Chapter 15 (Fats) however, is the only one for which the QUE estimator could not achieve a positive estimate for the variance of the time effect, contrary to the MINQUE procedure. For this reason, the corresponding variance was set to 0, and the value of the Hausman and LM test statistics using GLS-QUE are computed using this value for σ_3^2. Consequently, results for chapter 15 are better interpreted in terms of the GLS-MINQUE estimator alone, in particular when joint significance tests of variance components are concerned.

When considering MINQUE variance component estimates only, this has two important consequences: first, our variance components can be considered consistent, as well as the associated LM test statistics. Second, unobserved heterogeneity specific to products and countries do not appear to be correlated with prices and total expenditure ($\log \mathrm{EXP}_t$). Hence, while we can have some confidence in the inference we conduct about the structure of the multi-way error specification based on GLS estimates, we

[3] The implementation of QUE and MINQUE estimators was performed using the procedures provided by Peter Davis for the Gauss software, and described in Davis (2002).

Table 8.3. European import share equation – Meat

Parameter	OLS	Within	GLS-QUE	GLS-MINQUE
Chapter 2. Meat				
log price	−0.0563***	−0.0062	−0.0140	−0.0165
	(0.0128)	(0.0162)	(0.0153)	(0.0153)
log PEU	0.0032	−0.0163	−0.0135	−0.0120
	(0.0150)	(0.0165)	(0.0153)	(0.0154)
log PAR	0.0042	−0.0035	−0.0013	−0.0012
	(0.0052)	(0.0054)	(0.0052)	(0.0053)
log PAM	0.0128	0.0313**	0.0226*	0.0230*
	(0.0115)	(0.0132)	(0.0117)	(0.0119)
log EXP	−0.0002	−0.0003	−0.0010	−0.0008
	(0.0247)	(0.0252)	(0.0240)	(0.0243)
Intercept	0.1448	–	0.1385	0.1386
	(0.3424)		(0.3354)	(0.3385)
RMSE	0.1372	0.1439	0.1415	0.1410
Hausman test			7.6888	11.0448
			(0.1742)	(0.0505)
σ_1 (product)			0.0217	0.0227
σ_2 (country)			0.0655	0.0541
σ_3 (year)			0.0449	0.0439
σ_ε			0.1232	0.1249
LM test (i) σ_1 (product) $= 0$			4.2628	3.5301
			(0.0389)	(0.0602)
(ii) σ_2 (country) $= 0$			4.5827	4.3316
			(0.0323)	(0.0374)
(iii) σ_3 (year) $= 0$			5.3689	4.7569
			(0.0204)	(0.0291)
(i) + (ii)			6.3942	6.0168
			(0.0408)	(0.0493)
(i) + (iii)			3.7269	2.4552
			(0.1551)	(0.2929)
(ii) + (iii)			8.9170	8.4641
			(0.0115)	(0.0145)
(i) + (ii) + (iii)			9.5569	9.1379
			(0.0227)	(0.0275)
Observations	212			

Note. Standard errors and p-values are in parentheses for parameter estimates and test statistics respectively.

*10 percent level.

**5 percent level.

***1 percent level.

Table 8.4. European import share equation – Dairy products

Parameter	OLS	Within	GLS-QUE	GLS-MINQUE
Chapter 4. Dairy products				
log price	0.0073	−0.0287***	−0.0268***	−0.0268***
	(0.0067)	(0.0071)	(0.0069)	(0.0070)
log PEU	0.0079	0.0392***	0.0347***	0.0351***
	(0.0067)	(0.0080)	(0.0075)	(0.0077)
log PAR	−0.0163***	−0.0127***	−0.0115***	−0.0117***
	(0.0035)	(0.0035)	(0.0034)	(0.0034)
log PAM	0.0092*	0.0109**	0.0123**	0.0119**
	(0.0056)	(0.0051)	(0.0050)	(0.0050)
log EXP	0.0041	0.0034	0.0010	0.0014
	(0.0087)	(0.0084)	(0.0080)	(0.0082)
Intercept	−0.0128	–	−0.0170	−0.0243
	(0.1311)		(0.1221)	(0.1238)
RMSE	0.1180	0.1200	0.1197	0.1198
Hausman test			6.8675	7.7951
			(0.2306)	(0.1679)
σ_1 (product)			0.0368	0.0446
σ_2 (country)			0.0498	0.0449
σ_3 (year)			0.0136	0.0169
σ_ε			0.1009	0.1015
LM test (i) σ_1 (product) $= 0$			41.3105	31.5192
			(0.0000)	(0.0000)
(ii) σ_2 (country) $= 0$			46.1936	48.9460
			(0.0000)	(0.0000)
(iii) σ_3 (year) $= 0$			5.0716	4.1831
			(0.0243)	(0.0408)
(i) + (ii)			69.7937	69.8706
			(0.0000)	(0.0000)
(i) + (iii)			39.4286	36.0447
			(0.0000)	(0.0000)
(ii) + (iii)			37.4989	37.2523
			(0.0000)	(0.0000)
(i) + (ii) + (iii)			67.0836	65.8316
			(0.0000)	(0.0000)
Observations	902			

Note. Standard errors and *p*-values are in parentheses for parameter estimates and test statistics respectively.

*10 percent level.

**5 percent level.

***1 percent level.

Table 8.5. European import share equation – Cereals

Parameter	OLS	Within	GLS-QUE	GLS-MINQUE
Chapter 10. Cereals				
log price	−0.1688***	−0.0573***	−0.0608***	−0.0676***
	(0.0174)	(0.0179)	(0.0171)	(0.0170)
log PEU	0.0249	−0.0196	−0.0200	−0.0204
	(0.0202)	(0.0148)	(0.0142)	(0.0147)
log PAR	0.0419*	0.0571**	0.0508**	0.0561***
	(0.0236)	(0.0215)	(0.0197)	(0.0200)
log PAM	0.1157***	0.0288	0.0380*	0.0390**
	(0.0245)	(0.0206)	(0.0194)	(0.0198)
log EXP	0.0527	0.1386***	0.0956**	0.1219**
	(0.0604)	(0.0497)	(0.0422)	(0.0473)
Intercept	−0.4719	–	−1.2191**	−1.5765**
	(0.8518)		(0.6096)	(0.6741)
RMSE	0.3052	0.3359	0.3324	0.3307
Hausman test			7.0360	9.0991
			(0.2179)	(0.1051)
σ_1 (product)			0.1697	0.1005
σ_2 (country)			0.2864	0.1847
σ_3 (year)			0.0208	0.0558
σ_ε			0.1962	0.2002
LM test (i) σ_1 (product) $= 0$			12.2747	3.8388
			(0.0000)	(0.0500)
(ii) σ_2 (country) $= 0$			201.8000	59.0327
			(0.000)	(0.0000)
(iii) σ_3 (year) $= 0$			7.1331	2.6079
			(0.0075)	(0.1063)
(i) + (ii)			84.5920	76.1020
			(0.0000)	(0.0000)
(i) + (iii)			14.9273	10.6499
			(0.0005)	(0.0048)
(ii) + (iii)			66.8251	59.6952
			(0.0000)	(0.0000)
(i) + (ii) + (iii)			86.3018	82.2103
			(0.0000)	(0.0000)
Observations	261			

Note. Standard errors and *p*-values are in parentheses for parameter estimates and test statistics respectively.

*10 percent level.

**5 percent level.

***1 percent level.

Table 8.6. European import share equation – Animal and vegetable fats

Parameter	OLS	Within	GLS-QUE	GLS-MINQUE
Chapter 15. Animal and vegetable fats				
log price	−0.0620***	−0.0377***	−0.0364***	−0.0387***
	(0.0072)	(0.0065)	(0.0064)	(0.0064)
log PEU	−0.0031	0.0050	0.0035	0.0032
	(0.0060)	(0.0051)	(0.0050)	(0.0050)
log PAR	0.0138**	0.0152***	0.0148***	0.0159***
	(0.0058)	(0.0047)	(0.0046)	(0.0047)
log PAM	0.0212***	0.0009	−0.0030	0.0159
	(0.0071)	(0.0060)	(0.0058)	(0.0047)
log EXP	0.0171	0.0095	0.0070	0.0094
	(0.0108)	(0.0093)	(0.0084)	(0.0091)
Intercept	−0.1726	–	−0.0022	−0.0352
	(0.1555)		(0.1266)	(0.1385)
RMSE	0.1379	0.1397	0.1399	0.1396
Hausman test			14.3542	5.0533
			(0.0135)	(0.4094)
σ_1 (product)			0.0431	0.0277
σ_2 (country)			0.0942	0.1293
σ_3 (year)			0.0000	0.0141
σ_ε			0.1068	0.1076
LM test (i) σ_1 (product) $= 0$			8.1188	5.2606
			(0.0043)	(0.0218)
(ii) σ_2 (country) $= 0$			86.3235	78.2605
			(0.0000)	(0.0000)
(iii) σ_3 (year) $= 0$			1.9904	2.8631
			(0.1583)	(0.0906)
(i) + (ii)			111.9229	107.1756
			(0.0000)	(0.0000)
(i) + (iii)			8.3276	6.7249
			(0.0155)	(0.0346)
(ii) + (iii)			85.5278	82.4316
			(0.0000)	(0.0000)
(i) + (ii) + (iii)			110.7229	106.9949
			(0.0000)	(0.0000)
Observations	912			

Note. Standard errors and p-values are in parentheses for parameter estimates and test statistics respectively. The variance of time effects σ_3^2 was set to 0 in the QUE case, as no positive estimate was found.

**5 percent level.

***1 percent level.

Table 8.7. European import share equation – Sugar

Parameter	OLS	Within	GLS-QUE	GLS-MINQUE
Chapter 17. Sugar				
log price	-0.0088^{***}	-0.0058^{**}	-0.0066^{**}	-0.0067^{**}
	(0.0026)	(0.0028)	(0.0027)	(0.0027)
log PEU	0.0078^{**}	0.0057	0.0054	0.0054
	(0.0032)	(0.0036)	(0.0035)	(0.0035)
log PAR	−0.0017	0.0013	0.0012	0.0013
	(0.0033)	(0.0034)	(0.0033)	(0.0034)
log PAM	0.0029	0.0006	0.0016	0.0016
	(0.0034)	(0.0035)	(0.0034)	(0.0034)
log EXP	-0.0103^{**}	-0.0104^{*}	-0.0115^{**}	-0.0116^{**}
	(0.0053)	(0.0054)	(0.0053)	(0.0053)
Intercept	0.2332^{***}	–	0.1862^{**}	0.1852^{**}
	(0.0760)		(0.0771)	(0.0779)
RMSE	0.0745	0.0746	0.0746	0.0746
Hausman test			3.3943	5.4554
			(0.6394)	(0.3628)
σ_1 (product)			0.0211	0.0251
σ_2 (country)			0.0156	0.0209
σ_3 (year)			0.0115	0.0115
σ_ε			0.0692	0.0696
LM test (i) σ_1 (product) $= 0$			10.8080	10.6233
			(0.0010)	(0.0011)
(ii) σ_2 (country) $= 0$			12.2419	12.2729
			(0.0004)	(0.0004)
(iii) σ_3 (year) $= 0$			3.8765	3.0372
			(0.0489)	(0.0813)
(i) + (ii)			89.2117	88.4433
			(0.0000)	(0.0000)
(i) + (iii)			31.2294	29.1991
			(0.0000)	(0.0000)
(ii) + (iii)			53.4133	52.3849
			(0.0000)	(0.0000)
(i) + (ii) + (iii)			25.8925	25.4420
			(0.0000)	(0.0000)
Observations	846			

Note. Standard errors and *p*-values are in parentheses for parameter estimates and test statistics respectively.

*10 percent level.

**5 percent level.

***1 percent level.

can also conclude that unobserved characteristics of the products are not correlated with observed prices.

A possible interpretation of this result is that prices already convey all the necessary information on product characteristics, so that unobserved heterogeneity in import demand shares is orthogonal to the price level. Hence, although demand share for an imported good from country h can be systematically larger than the one associated with the same good exported from country k whatever the country-wise price level, the ranking of these countries in terms of demand shares does not depend on price. The same situation would also be true for two products i and j exported from the same country, h.

As far as variance components are concerned, it can be seen from Tables 8.3 to 8.7 that product and country effects are always statistically significant when GLS-QUE is used. With MINQUE variance components however, the test for nullity of the variance of product effects has a *p-value* slightly above the 5 percent level for Meat (Table 8.3) and Cereals (Table 8.5). Regarding time effects, as mentioned above, the QUE variance of λ for chapter 15 (Fats) was set to 0, as no positive estimate was found. The variance of time effects is not significantly different from 0 for Cereals, Fats and Sugar with MINQUE, and Fats with QUE variance components estimates. Chapter 2 (Meat) is the only case for which the joint test of a zero variance for both product and year does not reject the null at the 5 percent level, with either QUE or MINQUE estimates. In every other case, joint significance tests strongly reject the fact that a pair of variances (or the three variance components) is equal to 0.

Let us now turn to parameter estimates implied by the different methods. Estimation results for the Meat sub-sample (Table 8.3) are rather poor, with only 4 significant parameter estimates: own-price with OLS, log PAM with Within, GLS-QUE and GLS-MINQUE. In all 5 cases, OLS estimates of own-price coefficient ($\log \mathit{price}$) are always larger in absolute value than Within and GLS. The coefficient associated to expenditure ($\log \mathrm{EXP}_t$) is larger in absolute value when estimated with OLS in the case of Milk and Dairy products, and Fats, but is lower in the 3 other cases. When comparing Within and GLS estimates of own-price and cross-price parameters, there is no clear pattern either. For example, own-price coefficient Within estimates are higher in absolute value than their GLS counterparts in the case of Milk and Dairy products (Table 8.4, but are lower in absolute value in the other cases (Meat, Cereals, Fats and Sugar). Within cross-price parameter estimates are higher in absolute value than GLS in the case of Meat, Milk and Dairy (with the exception of $\log \mathrm{PAM}$) and Sugar (with the exception of $\log \mathrm{PAM}$). Whether QUE or MINQUE variance component estimates are used can lead to significant differences in the magnitude of slope parameter estimates, but only in a limited number of cases (at least when parameters are significantly different from 0). This is particularly true for the coefficient on $\log \text{EXP}_t$ in the case of Cereals (Table 8.5), $\log \text{PAR}$ in the case of Fats (Table 8.6).

Inspecting Tables 8.3 to 8.7, one can note that substitution effects as identified by significant cross-price coefficients are not always present, depending on the category of products and the exporting region. For Meat (chapter 2, Table 8.3), a significant cross-price effect exists only with American countries. Meat and Dairy products (chapter 4, Table 8.4) are characterized by significant and positive cross-price terms with European and American competitors, and negative cross-price effects with Arab and Regional countries. As far as Cereals are concerned (chapter 10, Table 8.5), the cross-price coefficients are positive and significant for Arab and Regional, and American exporters. For Fats (chapter 15, Table 8.6), positive and significant cross-effects with Arab and Regional countries only indicate some degree of complementarity with European exporters. Finally, the Sugar category (chapter 17, Table 8.7) is not associated to any degree of neither substitutability nor complementarity with any exporting region (or European competitors).

An interesting aspect of our demand model is the fact that cross-price effects associated to European competitors can be identified. The only category of products where significant parameters are found when estimated with Within and GLS is Milk and Dairy (chapter 4, Table 8.4). Parameter estimates associated to $\log P_{\text{EU}}$ are all positive and close to each other in magnitude, indicating a significant degree of substitution between European exporters to Lebanon. The reason for this may be that dairy products in particular are highly specific in terms of national image, and it is also a sector where products are expected to be more differentiated than, say, cereals, sugar or fats.

On the whole, estimates associated with own- and cross-prices are inferior to those found in Andayani and Tilley (1997), indicating a somewhat smaller sensitivity of import shares. This may be due to the level of disaggregation in our data, compared to their study on similar products (fruits) where time series were used. Moreover, that European, AM and AR competitors' prices should have positive coefficients as indicating substitution possibilities is not always verified from our estimates. However, a more detailed inspection of country-by-country export patterns would be needed in this respect. For example, France and Italy may have different specialization strategies regarding corn, wheat and rice.

As far as expenditure in agricultural products is concerned, the associated coefficient is significant only for chapters 10 (Cereals) and 17 (Sugar). Surprisingly, it does have neither the same sign nor a similar magnitude, indicating that the “expenditure effect” associated with cereals exported from Europe is positive, while it is negative for sugar products. Caution must be paid however when interpreting these results from the consumer point of view, as expenditure here concerns only the associated chapter, not total expenditure. It is therefore not relevant to interpret this effect as a pure income effect, but rather, as the impact of a change in the expenditure devoted to this particular chapter on European country-wise import shares. This point is also valid of course when interpreting elasticities, see below.

Concerning efficiency, the loss in efficiency of using Within instead of GLS estimates is not very important, and this is also true of OLS. Furthermore, GLS-QUE and GLS-MINQUE are rather close in terms of the magnitude of parameter estimates, although the difference in variance component estimates leads to different conclusions for the Hausman and LM test in a limited number of cases.

We finally compute Marshallian (uncompensated) elasticities of substitution between regions (EU, AR, AM and ROW), based on GLS-MINQUE estimates. For the case of European competitors, this elasticity is to be interpreted as a *“within-Europe”* substitution pattern. Since we only estimate a single share equation instead of the full system of import shares, we are only able to obtain a picture from one side of the market, the European one. Hence, it is not possible to infer substitution patterns between, say, Arab and Regional imports in the one hand, and American imports on the other. Results are given in Table 8.8, where we report elasticity estimates with their standard errors, from expressions given above in the section on the AIDS demand model.

Own-price elasticities are between -1 and -2 for all 5 commodity groups. In fact, a Student test for equality to -1.00 of this elasticity does not reject the null only in the case of the Cereals category. These figures are on average slightly lower than own-price elasticities found in Andayani and Tilley (1997) on a dataset of similarly disaggregated imports. They are not significantly different from the bilateral ones reported in Marquez (1990, 1994), and obtained under a variety of estimation procedures. Expenditure elasticities are all positive and significant, and close to 1 with the exception of Cereals (close to 2.00) and Sugar (0.64). This result might seem surprising at first, as it would indicate a much stronger reaction of import demand for cereals than for sugar from Europe when expenditure increases. As indicated above, expenditure is defined as discounted total demand (in value) for all products within the considered category (chapter). Hence, because of this separability restriction, the above effect should not be understood as a pure income effect, but instead as merely indicating that European cereals benefit very strongly from an increase in total demand for cereals, compared to other exporting regions.

Table 8.8. Marshallian demand elasticities

	Own-price	EU	AR	AM	ROW	Expenditure
Meat	−1.1609***	−0.1139***	−0.0116***	0.2248***	0.0795***	0.9919***
	(0.0354)	(0.0245)	(0.0025)	(0.0492)	(0.0177)	(0.0017)
Milk, Dairy	−1.6251***	0.7940**	−0.2737**	0.2773**	0.0687**	1.0338***
	(0.2906)	(0.3708)	(0.1279)	(0.1290)	(0.0337)	(0.0157)
Cereals	−1.7520***	−0.6213	0.3849	0.2254	0.8559	2.1365**
	(0.4638)	(0.5012)	(0.3752)	(0.2931)	(0.6907)	(0.8365)
Fats	−1.4385***	−0.0715*	0.1428**	−0.0362	0.0964**	1.1044***
	(0.1625)	(0.0383)	(0.0633)	(0.0336)	(0.0418)	(0.0395)
Sugar	−1.1902***	0.3281**	0.1236	0.1328*	−0.0382	0.6477***
	(0.0358)	(0.1229)	(0.0811)	(0.0814)	(0.0732)	(0.0624)

Note. Elasticities are based on GLS with MINQUE variance estimates. Own and Expenditure are own-price and expenditure elasticities respectively. EU, AR, AM and ROW respectively indicate elasticities of substitution between single-country European import price and European, Arab and Regional, American, and Rest of the World competitors.

*10 percent level.

**5 percent level.

***1 percent level.

Considering now cross-price elasticities, it can be seen that a majority of them are positive, with the exception of European competitors for Cereals (not significant) and Fats, Arab and Regional for Milk and Dairy, American exports for Fats (not significant), and Rest of the World for Sugar (not significant). When they are positive and significant, cross-price elasticities range between 8 and 79 percent, while only four are negative and significant: EU and AR for Meat, AR for Milk and Dairy, and EU for Fats. Hence, European imports appear substitutes for Milk and Dairy, and Sugar, while they are complementary for Meat and Fats. Arab and Regional country imports are significant substitutes to European imports for Fats only, and complementary to Meat, and Milk and Dairy. American imports are significant substitutes to European imports in the case of Meat, Milk and Dairy, and Sugar, while imports from the Rest of the World countries are significant substitutes to Meat, Milk and Dairy, and Fats.

8.6. Conclusion

This paper revisits the issue of estimating import elasticities, in the perspective of bilateral or multilateral trade agreements between countries and trade regions. While most empirical applications have used aggregate data in the form of time series in order to predict diversion and substitution patterns in international trade, we use for the first time to our knowledge, data directly obtained from a national customs administration for a single country (Lebanon). We are thus able to perform an empirical analysis of import shares at a very disaggregate level and with a much larger number of observations.

Since bilateral relationships between trade regions (regional blocks) are still essential in the current trend toward trade liberalization, it is interesting to address the issue of unobserved components in international commodity transactions, possibly not related to prices. As an example, quality of imported products, packaging standards but also implicit contracts between countries may explain a significant share of trade relations.

In order to model such unobserved trade factors, we first specify a simple micro-economic model of import share determination, using the AIDS (Almost Ideal Demand System) specification from the applied demand analysis. Restricting our attention to major agricultural commodities (meat, dairy products, cereals, animal and vegetable fats, sugar), we estimate an import share equation for European products as a function of own-price and competitors prices. Competition is taking place between European countries, Arab and Regional countries, North and South America, and the Rest of the World.

The econometric model incorporates a multi-way error components structure for unbalanced panel data, to accommodate for a more general heterogeneity pattern. Product, country, and time effects constitute separate unobserved effects whose influence is controlled for by panel data techniques (either fixed effects for conditional inference, or random effects for unconditional one). We estimate the import share equation by allowing parameter heterogeneity across the 5 commodity groups, and test for the validity of our multi-way error components specification with unbalanced data. Estimation results show that our specification is generally supported by the data, and that a more general error structure exists than what is generally considered in the literature, i.e., including country effects in addition to unobserved product effects. We also test for the random effects specification and do not reject it in favor of the fixed effects model, indicating that no significant correlation exists between product and country effects on the one hand, and import prices on the other. This last finding might be related to the fact that we only concentrate on agricultural commodities which are mostly homogeneous in nature. An interesting extension of our present approach would be to test our empirical specification on manufacturing goods that are highly differentiated (such as cars for instance).

Acknowledgements

The authors would like to thank the Editor and an anonymous referee for helpful comments and suggestions. Discussions with Thierry Magnac and Akiko Suwa-Eisenmann are gratefully acknowledged. We also thank Peter Davis for having shared his source code written in Gauss.

References

Alston, J.M., Carter, C.A., Green, R., Pick, D. (1990), "Whither Armington trade models?", *American Journal of Agricultural Economics,* Vol. 72, pp. 455–467.

Andayani, S.R.M., Tilley, D.S. (1997), "Demand and competition among supply sources: the Indonesian fruit export market", *Agricultural and Applied Economics,* Vol. 29, pp. 278–290.

Antweiler, W. (2001), "Nested random effects estimation in unbalanced panel data", *Journal of Econometrics,* Vol. 101, pp. 295–313.

Baltagi, B.H., Chang, Y.J. (1994), "Incomplete panels: a comparative study of alternative estimators for the unbalanced one-way error component regression model", *Journal of Econometrics,* Vol. 62, pp. 67–89.

Baltagi, B.H., Song, S.H., Jung, B.C. (2001), "The unbalanced nested error component regression model", *Journal of Econometrics*, Vol. 101, pp. 357–381.

Baltagi, B.H., Egger, P., Pfaffermayr, M. (2003), "A generalized design for bilateral trade flow models", *Economics Letters,* Vol. 80, pp. 391–397.

Crozet, M., Erkel-Rousse, H. (2004), "Trade performances, product quality perceptions, and the estimation of trade price elasticities", *Review of International Economics,* Vol. 12, pp. 108–129.

Davis, P. (2002), "Estimating multi-way error components models with unbalanced data structures", *Journal of Econometrics,* Vol. 106, pp. 67–95.

Deaton, A., Muellbauer, J. (1980), "An almost ideal demand system", *American Economic Review,* Vol. 70, pp. 312–326.

Egger, P., Pfaffermayr, M. (2003), "The proper panel econometric specification of the gravity equation: a three-way model with bilateral interaction effects", *Empirical Economics*, Vol. 28, pp. 571–580.

Fuller, W.A., Battese, G.E. (1973), "Transformations for estimation of linear models with nested error structure", *Journal of the American Statistical Association,* Vol. 68, pp. 626–632.

Green, R., Alston, J.M. (1990), "Elasticities in AIDS models", *American Journal of Agricultural Economics,* Vol. 72, pp. 442–445.

Hayes, D.J., Wahl, T.I., Williams, G.W. (1990), "Testing restrictions on a model of Japanese meat demand", *American Journal of Agricultural Economics,* Vol. 72, pp. 556–566.

Marquez, J. (1990), "Bilateral trade elasticities", *The Review of Economics and Statistics,* Vol. 72, pp. 70–77.

Marquez, J. (1994), "The econometrics of elasticities or the elasticity of econometrics: an empirical analysis of the behavior of U.S. imports", *The Review of Economics and Statistics,* Vol. 76, pp. 471–481.

Panagariya, A. (2000), “Preferential trade liberalization: the traditional theory and new developments”, *Journal of Economic Literature*, Vol. 35, pp. 287–331.

Panagariya, A., Shah, S., Mishra, D. (1996), “Demand elasticities in international trade: are they really low?”, University of Maryland working paper.

Rao, R.C. (1971), “Estimation of variance and covariance components-MINQUE theory”, *Journal of Multivariate Analysis*, Vol. 1, pp. 257–275.

Satyanarayana, V., Wilson, W.W., Johnson, D.D. (1997), “Import demand for malt: a time series and econometric analysis”, North Dakota State University Agricultural Economics Report # 349.

Searle, S.R. (1987), *Linear Models for Unbalanced Data*, Wiley, New York.

Wansbeek, T., Kapteyn, A. (1989), “Estimation of the error-components model with incomplete panel”, *Journal of Econometrics*, Vol. 41, pp. 341–361.

Winters, L.A. (2004), “Trade liberalization and economic performance: an overview”, *The Economic Journal*, Vol. 114, pp. 4–21.

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74009-3

CHAPTER 9

Can Random Coefficient Cobb–Douglas Production Functions be Aggregated to Similar Macro Functions?

Erik Biørn*,[a,b], Terje Skjerpen[b] and Knut R. Wangen[b]

[a]Department of Economics, University of Oslo, P.O. Box 1095, Blindern, 0317 Oslo, Norway
E-mail address: erik.biorn@econ.uio.no
[b]Research Department, Statistics Norway, P.O. Box 8131 Dep, 0033 Oslo, Norway
E-mail addresses: terje.skjerpen@ssb.no; knut.reidar.wangen@ssb.no

Abstract

Parametric aggregation of heterogeneous micro production technologies is discussed. A four-factor Cobb–Douglas function with normally distributed firm specific coefficient vector and with log-normal input vector (which agrees well with the available data) is specified. Since, if the number of micro units is large enough, aggregates expressed as arithmetic means can be associated with expectations, we consider conditions ensuring an approximate relation of Cobb–Douglas form to exist between expected output and expected inputs. Similar relations in higher-order moments also exist. It is shown how the aggregate input and scale elasticities depend on the coefficient heterogeneity and the covariance matrix of the log-input vector and hence vary over time. An implementation based on firm-level panel data for two manufacturing industries gives estimates of industry-level input elasticities and decomposition for expected output. Finally, aggregation biases when the correct aggregate elasticities are replaced by the expected firm-level elasticities, are explored.

Keywords: productivity, panel data, random coefficients, log-normal distribution, aggregate production function

JEL classifications: C23, C43, D21, L11

[*] Corresponding author.

9.1. Introduction

The production function is usually considered an essentially micro construct, and the existence, interpretation, and stability of a corresponding aggregate function are issues of considerable interest in macro-economic modeling and research, cf. the following quotations: "The benefits of an aggregate production model must be weighted against the costs of departures from the highly restrictive assumptions that underly the existence of an aggregate production function" (Jorgenson, 1995, p. 76) and "An aggregate production function is a function that maps aggregate inputs into aggregate output. But what exactly does this mean? Such a concept has been implicit in macroeconomic analyzes for a long time. However, it has always been plagued by conceptual confusions, in particular as to the link between the underlying micro production functions and the aggregate macro production function, *the latter thought to summarize the alleged aggregate technology*" (Felipe and Fisher, 2003, p. 209, our italics).[1] Four somewhat related questions are of interest: (Q1) Do the assumptions made ensure the *existence* of an aggregate function of the same parametric form as the assumed micro functions, and which additional assumptions will be required? (Q2) If the answer to (Q1) is in the affirmative, can a *naïve aggregation*, simply inserting mean values of micro parameters into a macro function with the same functional form, give an adequate representation of the 'aggregate technology'? (Q3) If the answer to (Q2) is in the negative, which are the most important *sources of aggregation bias* and instability of the parameters of the correctly aggregated macro function over time? (Q4) Does the heterogeneity of the micro technologies and/or the dispersion of the inputs across firms affect the macro parameters, and if so, how? Obviously, (Q4) is a following-up of (Q3).

Our focus in this paper will be on the four questions raised above, and we use a rather restrictive parametric specification of the *average* micro technology, based on a four-factor Cobb–Douglas function, with *random coefficients* to represent technological heterogeneity. Panel data is a necessity to examine such issues empirically in some depth. Yet, the intersection between the literature on aggregation and the literature on panel data econometrics is still small. Our study is intended to contribute both to methodological aspects and to give some new empirical evidence on the interface between linear aggregation of non-linear relations with parameter heterogeneity and panel data analysis. Although Cobb–Douglas

[1] A textbook exposition of theoretical properties of production functions aggregated from neo-classical micro functions is given in Mas-Colell *et al.* (1995, Section 5.E).

restricts input substitution strongly and has to some extent been rejected in statistical tests, the simplicity of this parametric form of the average technology is, for some applications, a distinctive advantage over, e.g., Translog or CES. In the empirical part of the paper, our focus will be on scale properties of the production technology, which can be well captured by the Cobb–Douglas form. We assume that the random coefficients are jointly normal (Gaussian) and that the inputs are generated by a multivariate log-normal distribution, whose parameters may shift over time. To our knowledge, this is the first study exploring aggregate production functions by using *firm-level* (unbalanced) panel data in a random coefficient setting by means of this form of the average micro technology.

A model framework which is similar to ours, although denoted as ‘cross-sectional aggregation of log-linear models’, is considered by van Garderen *et al.* (2000, Section 4.2) cf. also Lewbel (1990, 1992). However, on the one hand, we generalize some of their theoretical results to hold not only for first-order, but also for higher-order moments, on the other hand they illustrate the theoretical results, not on data from single firms, but on time series data from selected industries (p. 309), which is less consistent with the underlying micro theory. In our study the expectation vector and covariance matrix of the coefficient vector are estimated from panel data for two Norwegian manufacturing industries. Log-normality of the inputs is tested and for the most part not rejected. This, in conjunction with a Cobb–Douglas technology with normally distributed coefficients, allows us to derive interpretable expressions for the distribution of aggregate production.

From the general literature on aggregation it is known that properties of relationships aggregated from relationships for micro units, and more basically their existence, depend on the average functional form in the micro model, its heterogeneity, the distribution of the micro variables, and the form of the aggregation functions. A main concern in Stoker (1986b, 1993) is to warn against the use of representative agent models in macroeconometrics. The representative agent interpretation is valid only under rather restrictive assumptions, and in many realistic situations parameters in macro relations, which are analogous to those in the micro relations, cannot be given a structural interpretation since they represent a mixture of structural micro parameters and parameters characterizing the distribution of variables or parameters across the micro units. Furthermore, if these distributions vary over time, the correctly aggregated relation will be unstable, such that a constant parameter macro relation will be mis-specified and its application in time series contexts dubious. In our specific setting the micro parameters are not recoverable from macro data (cf. Stoker, 1993, p. 1843 for a discussion of this concept), so like Stoker (1986a) we are concerned with the opposite problem: how macroeconomic effects can be estimated by means of micro data. This goes some way in answering (Q1)–(Q4). Stoker only considers cross-sectional data for this purpose, but since random parameters, introduced to capture firm heterogeneity, are an integral part of our model, panel data are required to quantify macroeconomic parameters properly. Thus, in comparison with Stoker (1986a) the framework we consider is more data demanding.

Customarily, the aggregation functions are arithmetic means or sums. If the number of micro units is large enough to appeal to a statistical law of large numbers and certain additional statistical regularity conditions are satisfied, we can associate arithmetic means with expectations (cf. Fortin, 1991, Section 2; Stoker, 1993, Section 3; Hildenbrand, 1998, Section 2; Biørn and Skjerpen, 2004, Section 2), which is what we shall do here.

Given normality of both the coefficient vector and the log-input vector to which it belongs, output will not be log-normal marginally, and its distribution has to be examined specifically. We employ a formula for the expectation of output, which we generalize to also hold for higher-order origo moments of output, provided that they exist. The existence is guaranteed by an eigenvalue condition which involves the covariance matrices of the random coefficients and the log-inputs. Examining this condition for each year in the data period, we find that, generally, only the first- and second-order origo moments of output exist.

Besides the exact formulae, approximate expressions for the origo moments of output, which are easier to interpret, are considered. We find that the approximate formula performs fairly well for the first-order moment, whereas larger deviations occur for the second-order moment. Since aggregate parameters are in general undefined unless the distribution of the micro variables is restricted in some way, we provide results for the limiting cases where the means of the log-inputs change and their dispersions are preserved, and the opposite case. We denote these as the ‘mean preserving’ and ‘dispersion preserving’ parameters, respectively.

The maximal biases in scale elasticities brought about by comparing elasticities based on naïve analog formulae with those based on the dispersion preserving industry-level elasticities are about 9 and 7 per cent for Pulp and paper and Basic metals, respectively. Under the mean preserving definition the corresponding biases are about 26 and 15 per cent. Even larger biases are found for some of the input elasticities. Furthermore, we find differences in the ranking of the elasticities according to size when confronting correct industry-level elasticities and elasticities based on naïve analog formulae.

The rest of the paper is organized as follows. The model is presented in Section 9.2, properties of the theoretical distribution of output are discussed, and approximations to the moments of output are derived. From this we obtain, in Section 9.3, an approximate aggregate production function and expressions for the correct aggregate input elasticities, in the dispersion preserving and mean preserving cases. In Section 9.4, the data are described and the estimation of the micro structure is discussed, and then, in Section 9.5 estimates of aggregate input elasticities are presented and other applications indicated. Section 9.6 concludes.

9.2. Model and output distribution

9.2.1. Basic assumptions

We consider an n factor Cobb–Douglas production function model for panel data, written in log-linear form as

$$y_{it} = x'_{it}\beta_i + u_{it} = \alpha_i + z'_{it}\gamma_i + u_{it}, \tag{9.1}$$

where i is the firm index, t is the period index, $x_{it} = (1, z'_{it})'$ is an $n+1$ vector (including a one for the intercept) and $\beta_i = (\alpha_i, \gamma'_i)'$ is an $n+1$ vector (including the intercept), γ_i denoting the n vector of input elasticities and u_{it} is a disturbance. We interpret z_{it} as $\ln(Z_{it})$, where Z_{it} is the n-dimensional input vector, and y_{it} as $\ln(Y_{it})$, where Y_{it} is output, and assume that all β_i and u_{it} are stochastically independent and independent of all x_{it} and that

$$x_{it} \sim \mathcal{N}(\mu_{xt}, \Sigma_{xxt}) = \mathcal{N}\left(\begin{bmatrix} 1 \\ \mu_{zt} \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{zzt} \end{bmatrix}\right), \tag{9.2}$$

$$\beta_i \sim \mathcal{N}(\mu_\beta, \Sigma_{\beta\beta}) = \mathcal{N}\left(\begin{bmatrix} \mu_\alpha \\ \mu_\gamma \end{bmatrix}, \begin{bmatrix} \sigma_{\alpha\alpha} & \sigma'_{\gamma\alpha} \\ \sigma_{\gamma\alpha} & \Sigma_{\gamma\gamma} \end{bmatrix}\right), \tag{9.3}$$

$$u_{it} \sim \mathcal{N}(0, \sigma_{uu}), \tag{9.4}$$

where $\sigma_{\alpha\alpha} = \operatorname{var}(\alpha)$, $\sigma_{\gamma\alpha}$ is the n vector of covariances between α_i and γ_i, and $\Sigma_{\gamma\gamma}$ is the n-dimensional covariance matrix of γ_i. The $n+1$-dimensional covariance matrix Σ_{xxt} is singular since x_{it} has a one element, while its n-dimensional submatrix Σ_{zzt} is non-singular in general. In the econometric model version [see (9.29) below] the description of the technology is a bit more general, since α_i also includes a linear deterministic trend.

9.2.2. The conditional distribution of output

We first characterize the distribution of *log-output*. From (9.1)–(9.4) it follows that

$$(y_{it}|\beta_i) \sim \mathcal{N}(\mu'_{xt}\beta_i, \beta'_i\Sigma_{xxt}\beta_i + \sigma_{uu}),$$
$$(y_{it}|x_{it}) \sim \mathcal{N}(x'_{it}\mu_\beta, x'_{it}\Sigma_{\beta\beta}x_{it} + \sigma_{uu}), \tag{9.5}$$

and, by using the law of iterated expectations, that

$$\mu_{yt} = \mathsf{E}(y_{it}) = \mathsf{E}\big[\mathsf{E}(y_{it}|\beta_i)\big] = \mu'_{xt}\mu_\beta, \tag{9.6}$$

$$\begin{aligned}\sigma_{yyt} &= \text{var}(y_{it}) = \mathsf{E}\big[\text{var}(y_{it}|\beta_i)\big] + \text{var}\big[\mathsf{E}(y_{it}|\beta_i)\big]\\ &= \mathsf{E}\big[\text{tr}(\beta_i\beta'_i\Sigma_{xxt}) + \sigma_{uu}\big] + \text{var}(\mu'_{xt}\beta_i)\\ &= \text{tr}\big[\mathsf{E}(\beta_i\beta'_i\Sigma_{xxt})\big] + \sigma_{uu} + \mu'_{xt}\Sigma_{\beta\beta}\mu_{xt}\\ &= \text{tr}\big[(\mu_\beta\mu'_\beta + \Sigma_{\beta\beta})\Sigma_{xxt}\big] + \sigma_{uu} + \mu'_{xt}\Sigma_{\beta\beta}\mu_{xt}\\ &= \mu'_{xt}\Sigma_{\beta\beta}\mu_{xt} + \mu'_\beta\Sigma_{xxt}\mu_\beta + \text{tr}(\Sigma_{\beta\beta}\Sigma_{xxt}) + \sigma_{uu}.\end{aligned} \tag{9.7}$$

The four components of σ_{yyt} represent (i) the variation in the coefficients $(\mu'_{xt}\Sigma_{\beta\beta}\mu_{xt})$, (ii) the variation in the log-inputs $(\mu'_\beta\Sigma_{xxt}\mu_\beta)$, (iii) the interaction between the variation in the log-inputs and the coefficients $[\text{tr}(\Sigma_{\beta\beta}\Sigma_{xxt})]$, and (iv) the disturbance variation (σ_{uu}).

We next characterize the distribution of *output*. Since $Y_{it} = \mathrm{e}^{y_{it}} = \mathrm{e}^{x'_{it}\beta_i + u_{it}}$, we know from (9.5) that $(Y_{it}|x_{it})$ and $(Y_{it}|\beta_i)$ are log-normal. From Evans *et al.* (1993, Ch. 25) it therefore follows, for any positive integer r, that

$$\begin{aligned}\mathsf{E}\big(Y^r_{it}|\beta_i\big) &= \mathsf{E}_{x_{it},u_{it}}\big(\mathrm{e}^{ry_{it}}|\beta_i\big)\\ &= \exp\left[r\mu'_{xt}\beta_i + \frac{1}{2}r^2(\beta'_i\Sigma_{xxt}\beta_i + \sigma_{uu})\right],\end{aligned} \tag{9.8}$$

$$\begin{aligned}\mathsf{E}\big(Y^r_{it}|x_{it}\big) &= \mathsf{E}_{\beta_i,u_{it}}\big(\mathrm{e}^{ry_{it}}|x_{it}\big)\\ &= \exp\left[rx'_{it}\mu_\beta + \frac{1}{2}r^2(x'_{it}\Sigma_{\beta\beta}x_{it} + \sigma_{uu})\right],\end{aligned} \tag{9.9}$$

which show that any *conditional finite-order* moment of output when conditioning on the coefficient vector or on the input vector exists. These equations are interesting as far as they go, but we will also need the marginal moments of output.

9.2.3. Exact marginal origo moments of output

Assuming that the rth-order origo moment of Y_{it} exists, (9.8) and the law of iterated expectations yield

$$\mathsf{E}\big(Y^r\big) = \exp\left[\frac{1}{2}r^2\sigma_{uu}\right]\mathsf{E}_\beta\left[\exp\left(r\mu'_x\beta + \frac{1}{2}r^2\beta'_i\Sigma_{xx}\beta\right)\right]$$

$$= \exp\left[r\mu_x'\mu_\beta + \frac{1}{2}r^2(\mu_\beta'\Sigma_{xx}\mu_\beta + \sigma_{uu})\right]$$
$$\times \mathsf{E}_\delta\left\{\exp\left[\left(r\mu_x' + r^2\mu_\beta'\Sigma_{xx}\right)\delta + \frac{1}{2}r^2\delta'\Sigma_{xx}\delta\right]\right\}, \tag{9.10}$$

where $\delta = \beta - \mu_\beta \sim \mathcal{N}(0, \Sigma_{\beta\beta})$ and subscripts (i, t) are from now on omitted for simplicity. A closed form expression for the marginal rth-order origo moments can be derived from (9.10). In Appendix A9.1 it is shown that:[2]

$$\begin{aligned}&\mathsf{E}\left(Y^r\right)\\ &\quad = \left|I_{n+1} - r^2\Sigma_{\beta\beta}\Sigma_{xx}\right|^{-1/2}\exp\left[r\mu_x'\mu_\beta + \frac{1}{2}r^2(\mu_\beta'\Sigma_{xx}\mu_\beta + \sigma_{uu})\right.\\ &\qquad \left.+\frac{1}{2}\left(r\mu_x' + r^2\mu_\beta'\Sigma_{xx}\right)\left(\Sigma_{\beta\beta}^{-1} - r^2\Sigma_{xx}\right)^{-1}\left(r\mu_x + r^2\Sigma_{xx}\mu_\beta\right)\right].\end{aligned} \tag{9.11}$$

It is obvious from this expression that the existence of $\mathsf{E}(Y^r)$ requires that the inverse of $(\Sigma_{\beta\beta}^{-1} - r^2\Sigma_{xx})$ exists, but the derivations in Appendix A9.1, imply a stronger requirement:

$$\mathsf{E}\left(Y^r\right) \text{ exists } \iff \Sigma_{\beta\beta}^{-1} - r^2\Sigma_{xx} \text{ is positive definite.} \tag{9.12}$$

Observe that if $\mathsf{E}(Y^r)$ exists, all lower-order moments exist: let $M(r) = \Sigma_{\beta\beta}^{-1} - r^2\Sigma_{xx}$, then

$$M(r-1) = M(r) + (2r-1)\Sigma_{xx}, \quad r = 2, 3, \ldots. \tag{9.13}$$

If $M(r)$ and Σ_{xx} are positive definite, then $M(r-1)$ is also positive definite, since $2r > 1$ and the sum of two positive definite matrices is positive definite.

Paying attention to the existence of moments may seem less important than it actually is. The primary reason is that we will need approximation formulae to derive interpretable decompositions of the moments and expressions for the aggregate elasticities. Such approximations are only meaningful when the moments exist. A secondary reason is that if (9.2)–(9.4) had been replaced by other distributional assumptions, it would generally not be possible to express moments of output in closed form. In such cases, one could be tempted to estimate $\mathsf{E}(Y^r)$ by simulations. To illustrate: if the closed form expression (9.11) were unavailable, we could

[2] Following a similar argument, the same result can be derived from (9.9). In the case $r = 1$, a related and somewhat longer derivation based on an equation for the conditional expectation similar to (9.9) is provided by van Garderen *et al.* (2000, pp. 306–307).

have used simulations based on (9.10) and a large sample of synthetic β_i's drawn from a distribution given by (9.3). For each synthetic β_i, the right-hand side of (9.10) can be calculated, and moments estimated by averaging over the whole sample of β_i's. When the true moment exists, the law of large numbers ensures that estimates of such a procedure would converge towards the true moments as the sample size increases. However, the convergence may be extremely slow if certain higher-order moments required by the familiar standard central limit theorems do not exist. In such cases, a more general central limit theorem is appropriate, see Embrechts *et al.* (1997, pp. 71–81).[3] The importance of existence of higher-order moments in simulation-based estimation can be illustrated by the empirical application below where we find that the order of the highest existing moment is usually two. This means that simulations would work well for first-order moments, that very large simulation samples could be required to obtain precise estimates of second-order moments, and that simulated estimates of (non-existing) higher-order moments would be numerically unstable for any sample size.

9.2.4. *Approximations to the marginal origo moments of output*

Equation (9.11), although in closed form, cannot be easily interpreted and decomposed, mainly because of the determinant expression and the inverse covariance matrix $\Sigma_{\beta\beta}^{-1}$ which it contains. We now present a way of obtaining, from (9.10), an approximate formula for $\mathsf{E}(Y^r)$, which is simpler to interpret.

Provided that (9.12) holds, an approximation to the rth-order origo moment of output can be obtained by replacing $\delta'\Sigma_{xx}\delta = \operatorname{tr}[\delta\delta'\Sigma_{xx}]$ by its expected value, $\operatorname{tr}[\Sigma_{\beta\beta}\Sigma_{xx}]$, in the exponent in the argument of $\mathsf{E}_\delta\{\cdot\}$ in (9.10). Letting $a(r) = r\mu_x + r^2\Sigma_{xx}\mu_\beta$ we obtain

$$\begin{aligned}\mathsf{E}\big(Y^r\big) \approx G_r(Y) &= \exp\bigg[r\mu_x'\mu_\beta + \frac{1}{2}r^2\big(\mu_\beta'\Sigma_{xx}\mu_\beta \\ &\qquad + \operatorname{tr}[\Sigma_{\beta\beta}\Sigma_{xx}] + \sigma_{uu}\big)\bigg]\mathsf{E}\big[\exp\big(a(r)'\delta\big)\big] \\ &= \exp\bigg[r\mu_x'\mu_\beta + \frac{1}{2}r^2\big(\mu_\beta'\Sigma_{xx}\mu_\beta \\ &\qquad + \operatorname{tr}[\Sigma_{\beta\beta}\Sigma_{xx}] + \sigma_{uu}\big)\bigg]\exp\bigg[\frac{1}{2}a(r)'\Sigma_{\beta\beta}a(r)\bigg],\end{aligned}$$

[3] If, for instance, $x_1, \ldots, x_m$ denote a sequence of random variables with mean μ and variance σ^2, the sample average converges towards μ with a rate $1/\sqrt{m}$. But if the variance does not exist, the rate of convergence cannot be established by standard central limit theorems.

since $\delta \sim \mathcal{N}(0, \Sigma_{\beta\beta})$. Rearranging gives

$$\begin{aligned}\mathsf{E}(Y^r) \approx G_r(Y) = \exp\Big[& r\mu_x'\mu_\beta + \frac{1}{2}r^2\big(\mu_\beta'\Sigma_{xx}\mu_\beta + \mu_x'\Sigma_{\beta\beta}\mu_x \\ & + \operatorname{tr}[\Sigma_{\beta\beta}\Sigma_{xx}] + \sigma_{uu}\big) + r^3\mu_\beta'\Sigma_{xx}\Sigma_{\beta\beta}\mu_x \\ & + \frac{1}{2}r^4\mu_\beta'\Sigma_{xx}\Sigma_{\beta\beta}\Sigma_{xx}\mu_\beta\Big]. \end{aligned} \tag{9.14}$$

When (9.12) does not hold, this approximation, of course, makes no sense. When applying the approximation $G_r(Y)$ we eliminate both the square root of the inverse of the determinant $|I_{n+1} - r^2\Sigma_{\beta\beta}\Sigma_{xx}|$ and all terms involving $\Sigma_{\beta\beta}^{-1}$ from the function. This is an obvious simplification when we use it to derive *and, more importantly, interpret* expressions for the aggregate input and scale elasticities below.

Our intuition says that $G_r(Y)$ is likely to underestimate $\mathsf{E}(Y^r)$, since the main difference between them is that the former has a reduced spread in the exponent of the convex exponential function, compared to the latter. The argument is that when deriving (9.14), we neglect the dispersion of the quadratic form $\delta'\Sigma_{xx}\delta$, where δ has a symmetric distribution. This way of reasoning will be supported by the results in Section 9.5.1.

We can then, using (9.6) and (9.7), write the analytical approximation to $\mathsf{E}(Y^r)$ as

$$G_r(Y) = \Phi_r(y)\Gamma_r\Lambda_r, \tag{9.15}$$

where

$$\Phi_r(y) = \exp\left[r\mu_y + \frac{1}{2}r^2\sigma_{yy}\right] \tag{9.16}$$

is the *'first-order' approximation* we would have obtained if we had proceeded as if y were normally and Y were log-normally distributed marginally, and

$$\begin{aligned}\Gamma_r &= \exp\big[r^3\mu_x'\Sigma_{\beta\beta}\Sigma_{xx}\mu_\beta\big], \\ \Lambda_r &= \exp\left[\frac{1}{2}r^4\mu_\beta'\Sigma_{xx}\Sigma_{\beta\beta}\Sigma_{xx}\mu_\beta\right],\end{aligned} \tag{9.17}$$

where Λ_r and Γ_r can be considered *correction factors* which serve to improve the approximation. Note that the exponent in the expression for Λ_r is a positive definite quadratic form whenever $\Sigma_{\beta\beta}$ is positive definite, while the exponent in the expression for Γ_r can have either sign.[4] In the

[4] The origin of the approximation leading to (9.14) is (9.8). Proceeding in a similar way from (9.9) would have given a symmetric approximation, with a different Λ_r component; see Biørn *et al.* (2003b, Sections 3.1 and 6.3).

special case with no coefficient heterogeneity (i.e., $\Sigma_{\beta\beta}$ is a zero matrix), (9.8) and (9.14) give identical results, and then $\mathsf{E}(Y^r) = G_r(Y) = \Phi_r(y)$ and $\Gamma_r = \Lambda_r = 1$ for all r.

9.3. *An approximate aggregate production function in origo moments*

We next derive an approximate relationship between $\mathsf{E}(Y^r)$ and $\mathsf{E}(Z^r)$ to be used in examining aggregation biases when the aggregate variables are represented by their arithmetic means. In doing this, we recall that $\mathrm{e}^{\mathsf{E}[\ln(Y)]}$ and $\mathrm{e}^{\mathsf{E}[\ln(Z)]}$ correspond to the geometric means, and $\mathsf{E}(Y)$ and $\mathsf{E}(Z)$ to the arithmetic means of the output and the input vector, respectively. We first assume r arbitrarily large, still assuming that (9.12) is satisfied, and afterwards discuss the case $r = 1$ in more detail.

9.3.1. *A Cobb–Douglas production function in origo moments*

Let

$$\begin{aligned}\theta_{yr} &= \ln\big[G_r(Y)\big] - r\mu_y = \ln\big[\Phi_r(y)\big] + \ln[\Gamma_r] + \ln[\Lambda_r] - r\mu_x'\mu_\beta \\ &= \frac{1}{2}r^2\big[\mu_x'\Sigma_{\beta\beta}\mu_x + \mu_\beta'\Sigma_{xx}\mu_\beta + \mathrm{tr}(\Sigma_{\beta\beta}\Sigma_{xx}) + \sigma_{uu}\big] \\ &\quad + r^3\mu_x'\Sigma_{\beta\beta}\Sigma_{xx}\mu_\beta + \frac{1}{2}r^4\mu_\beta'\Sigma_{xx}\Sigma_{\beta\beta}\Sigma_{xx}\mu_\beta,\end{aligned} \tag{9.18}$$

after inserting from (9.7), which can be interpreted as an approximation to $\ln[\mathsf{E}(Y^r)] - \mathsf{E}[\ln(Y^r)]$. Let Z_j denote the jth element of the input vector Z and $z_j = \ln(Z_j)$. Since $z_j \sim \mathcal{N}(\mu_{zj}, \sigma_{zjzj})$, where μ_{zj} is the jth element of μ_z and σ_{zjzj} is the jth diagonal element of Σ_{zz} [cf. (9.2)], we have

$$\begin{aligned}&\mathsf{E}\big(Z_j^r\big) = \mathsf{E}\big(\mathrm{e}^{z_j r}\big) = \exp\left(\mu_{zj}r + \frac{1}{2}\sigma_{zjzj}r^2\right), \\ &\quad r = 1, 2, \ldots;\ j = 1, \ldots, n.\end{aligned} \tag{9.19}$$

Let $\mu_{\gamma j}$ be the jth element of μ_γ, i.e., the expected elasticity of the jth input. Since (9.19) implies $\exp(\mu_{zj}\mu_{\gamma j}r) = \exp(-\frac{1}{2}\sigma_{zjzj}r^2\mu_{\gamma j})[\mathsf{E}(Z_j^r)]^{\mu_{\gamma j}}$, it follows from (9.18) that

$$G_r(Y) = \mathrm{e}^{\mu_\alpha r}A_r\prod_{j=1}^{n}\big[\mathsf{E}\big(Z_j^r\big)\big]^{\mu_{\gamma j}}, \tag{9.20}$$

where

$$A_r = \exp\left(\theta_{yr} - \frac{1}{2}r^2\sum_{j=1}^{n}\sigma_{zjzj}\mu_{\gamma j}\right) = \exp\left(\theta_{yr} - \frac{1}{2}r^2\mu_\gamma'\sigma_{zz}\right), \tag{9.21}$$

and $\sigma_{zz} = \text{diagv}(\Sigma_{zz})$.[5] Equation (9.20) can be interpreted (approximately) as a *Cobb–Douglas function in the* r*th-order origo moments of* Y *and* $Z_1, \dots, Z_n$, with exponents equal to the expected firm-level elasticities $\mu_{\gamma 1}, \dots, \mu_{\gamma n}$ and an intercept $\mathrm{e}^{\mu_\alpha r}$, adjusted by the factor A_r. The latter depends, via θ_{yr}, on the first- and second-order moments of the log-input vector z and the coefficient vector β and σ_{uu}, cf. (9.18) and (9.21).

For $r = 1$, (9.20) gives

$$G_1(Y) = \mathrm{e}^{\mu_\alpha} A_1 \prod_{j=1}^{n} \left[\mathsf{E}(Z_j)\right]^{\mu_{\gamma j}}. \tag{9.22}$$

Seemingly, this equation could be interpreted as a Cobb–Douglas function in the arithmetic means $\mathsf{E}(Y)$ and $\mathsf{E}(Z_1), \dots, \mathsf{E}(Z_n)$, with elasticities coinciding with the expected firm-level elasticities $\mu_{\gamma 1}, \dots, \mu_{\gamma n}$ and an intercept e^{μ_α} adjusted by the factor A_1. In some sense one could then say that the aggregation problem had been “solved”. However, we will show that, due to the randomness of the micro coefficients in combination with the *non-linearity* of the micro function the situation is not so simple. As emphasized by Stoker (1993, p. 1846) introducing random coefficients in a *linear* equation will not bias the expectation of the endogenous variable; cf. also Zellner (1969).

9.3.2. Aggregation by analogy and aggregation biases in output and in input elasticities

Assume that we, instead of (9.22), represent the aggregate production function simply by

$$\widehat{\mathsf{E}(Y)} = \mathrm{e}^{\mu_\alpha} \prod_{j=1}^{n} \left[\mathsf{E}(Z_j)\right]^{\mu_{\gamma j}}. \tag{9.23}$$

This can be said to mimic the *aggregation by analogy*, or *naïve aggregation*, often used by macro-economists and macro model builders. The resulting *aggregation error in output*, when we approximate $\mathsf{E}(Y)$ by $G_1(Y)$, is

$$\varepsilon(Y) = G_1(Y) - \widehat{\mathsf{E}(Y)} = (A_1 - 1)\mathrm{e}^{\mu_\alpha} \prod_{j=1}^{n} \left[\mathsf{E}(Z_j)\right]^{\mu_{\gamma j}}. \tag{9.24}$$

Representing the aggregate Cobb–Douglas production function by (9.23) will bias not only its intercept, but also its derived input elasticities,

[5] We let ‘diagv’ before a square matrix denote the column vector containing its diagonal elements.

because A_1 in (9.22) is affected by changes in μ_z and Σ_{zz}. Equations (9.6), (9.7) and (9.18) show that when $\Sigma_{\gamma\gamma}$ is non-zero, a change in μ_z affects not only the expectation of log-output, μ_y, but also its variance σ_{yy}. Equations (9.15)–(9.17) imply

$$\ln\bigl[G_1(Y)\bigr] = \mu_y + \frac{1}{2}\sigma_{yy} + \mu_x' \Sigma_{\beta\beta} \Sigma_{xx} \mu_\beta + \frac{1}{2}\mu_\beta' \Sigma_{xx} \Sigma_{\beta\beta} \Sigma_{xx} \mu_\beta. \tag{9.25}$$

Using the fact that $\Delta \ln[\mathsf{E}(Z)] = \Delta(\mu_z + \frac{1}{2}\sigma_{zz})$ [cf. (9.19)], we show in Appendix A9.2 that

$$\frac{\partial \ln[G_1(Y)]}{\partial \ln[\mathsf{E}(Z)]} = \begin{cases} \mu_\gamma + \sigma_{\gamma\alpha} + \Sigma_{\gamma\gamma}(\mu_z + \Sigma_{zz}\mu_\gamma), & \text{when } \Sigma_{zz} \text{ is constant,} \\ \text{diagv}[\mu_\gamma\mu_\gamma' + \Sigma_{\gamma\gamma} + \mu_\gamma\mu_\gamma'\Sigma_{zz}\Sigma_{\gamma\gamma} & \text{when } \mu_z \text{ and the} \\ \qquad + \Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma\mu_\gamma' & \quad \text{off-diagonal elements} \\ \qquad + 2\mu_\gamma(\sigma_{\gamma\alpha}' + \mu_z'\Sigma_{\gamma\gamma})], & \quad \text{of } \Sigma_{zz} \text{ are constant.} \end{cases} \tag{9.26}$$

Hence we can not uniquely define and measure an exact aggregate jth input elasticity, $\partial \ln[G_1(Y)]/\partial \ln[\mathsf{E}(Z_j)]$ unless we restrict the way in which $\ln[\mathsf{E}(Z)]$ changes. The two parts of (9.26) are limiting cases, the first may be interpreted as a vector of *dispersion preserving* aggregate input elasticities, the second as a vector of *mean preserving* aggregate elasticities. Anyway, when $\Sigma_{\gamma\gamma}$ and $\sigma_{\gamma\alpha}$ are non-zero, μ_γ provides a biased measure of the aggregate elasticity vector. Dispersion preserving elasticities may be of more practical interest than mean preserving ones, since constancy of the *variance* of the log-input j, σ_{zjzj}, implies constancy of the *coefficient of variation* of the untransformed input j. This will be the situation when the jth input of all micro units change proportionally. This follows from the fact that the coefficient of variation of Z_j is (cf. (9.19) and Evans *et al.*, 1993, Ch. 25)

$$v(Z_j) = \frac{\text{std}(Z_j)}{\mathsf{E}(Z_j)} = \bigl(e^{\sigma_{zjzj}} - 1\bigr)^{1/2}, \tag{9.27}$$

and hence constancy of σ_{zjzj} implies constancy of $v(Z_j)$. Mean preserving elasticities relate to the more ‘artificial’ experiment with $\mathsf{E}[\ln(Z_j)]$ kept fixed and $v(Z_j)$ increased by increasing $\text{std}(Z_j)$. Our term dispersion preserving aggregation is related to the concept ‘mean scaling’ introduced by Lewbel (1990, 1992) in the context of aggregating log-linear relations.

The bias vector implied by the dispersion preserving aggregate input elasticities, obtained from the first part of (9.26), is

$$\varepsilon(\mu_\gamma) = \sigma_{\gamma\alpha} + \Sigma_{\gamma\gamma}(\mu_z + \Sigma_{zz}\mu_\gamma). \tag{9.28}$$

The bias vector for the mean preserving elasticities can be obtained from the second part in a similar way.

9.4. Data, microeconometric model and micro estimation

Unbalanced panel data sets for two manufacturing industries, Pulp and paper (2823 observations, 237 firms) and Basic metals (2078 observations, 166 firms) for the years 1972–1993 are used in the empirical application. These two export-oriented, energy-intensive industries are important for the Norwegian economy and accounted for almost one fourth of the mainland export-income in the sample period. Confronting our single-output multiple-input framework with data for these two industries also has the advantage that their outputs are rather homogeneous and hence can be measured in physical units rather than, e.g., deflated sales, which may be subject to measurement errors. A further description is given in Appendix B9.2.

Four inputs ($n = 4$) are specified: capital (K), labor (L), energy (E) and materials (M). A deterministic trend is intended to capture the level of the technology. We parameterize (9.1) as

$$y_{it} = \alpha_i^* + \kappa t + z_{it}'\gamma_i + u_{it}, \tag{9.29}$$

where $z_{it} = (z_{Kit}, z_{Lit}, z_{Eit}, z_{Mit})'$ is the log-input vector of firm j in period t. The parameter α_i^* and the parameter vector γ_i are random and specific to firm i, whereas κ is a firm invariant trend coefficient. With this change of notation, (9.3) reads $\psi_i \sim \mathcal{N}(\psi, \Omega)$, where $\psi = (\mu_\alpha^*, \mu_K, \mu_L, \mu_E, \mu_M)'$ and Ω are the expectation and the unrestricted variance–covariance matrix of the random parameters, respectively.

The unknown parameters are estimated by Maximum Likelihood (ML) using the PROC MIXED procedure in the SAS/STAT software (see Littell *et al.*, 1996) and imposing positive definiteness of Ω. In Appendix B9.1, the log-likelihood underlying estimation for our unbalanced panel data is formulated. This particular application relies on ML-estimation results in Biørn *et al.* (2002, cf. Section 2 and Appendix A, part 2). The estimates of ψ and κ and of the expected scale elasticity $\mu = \sum_j \mu_j$ are given in Table 9.1, whereas the estimate of Ω is given in Table 9.2. The estimated variances of the genuine error term are 0.0408 and 0.0986 for Pulp and paper and Basic metals, respectively.

Compared with the random intercept specification and with a model with no heterogeneity at all, our random coefficients model gives a substantially better *goodness of fit*. Going from the model with heterogeneity only in the intercept to the random coefficients model, the log-likelihood

***Table 9.1.* Firm-level Cobb–Douglas production functions. Parameter estimates**

	Pulp and paper		Basic metals	
	Estimate	St.err.	Estimate	St.err.
μ_α^*	−2.3021	0.2279	−3.1177	0.2702
κ	0.0065	0.0013	0.0214	0.0021
μ_K	0.2503	0.0344	0.1246	0.0472
μ_L	0.1717	0.0381	0.2749	0.0550
μ_E	0.0854	0.0169	0.2138	0.0374
μ_M	0.5666	0.0309	0.4928	0.0406
μ	1.0740	0.0287	1.1061	0.0324

***Table 9.2.* Firm-level Cobb–Douglas production functions. Covariance matrix of firm specific coefficients. Variances on the diagonal, correlation coefficients below**

	α^*	γ_K	γ_L	γ_E	γ_M
Pulp and paper					
α^*	5.9336				
γ_K	−0.4512	0.1147			
γ_L	−0.7274	−0.0559	0.1515		
γ_E	0.3968	−0.4197	−0.3009	0.0232	
γ_M	0.3851	−0.6029	−0.4262	0.1437	0.1053
Basic metals					
α^*	3.5973				
γ_K	−0.0787	0.1604			
γ_L	−0.6846	−0.5503	0.1817		
γ_E	0.3040	−0.6281	0.1366	0.1190	
γ_M	0.1573	0.1092	−0.3720	−0.6122	0.1200

value increases by about 365 and 200 in Pulp and paper and Basic metals, respectively, while the increase in the number of parameters is only 14. The corresponding increases when comparing our random coefficients model with a model without any firm-specific heterogeneity are 2045 and 1572, with an increase in the number of parameters of 15.

The implied expected scale elasticities are 1.07 in Pulp and paper and 1.11 in Basic metals (Table 9.1), indicating weak economies of scale; size and ranking of the expected input elasticities differ somewhat more. The estimates of the trend coefficients indicate that technical progress has been stronger in Basic metals than in Pulp and paper, 2.1 per cent and 0.6 per cent, respectively. As can be seen from the off-diagonal elements in Table 9.2 the pattern of correlation of the input elasticities across firms are somewhat different in the two industries, also with respect to sign.

Whereas normality of the log-input vector is not needed when estimating the micro structure, since the log-likelihood is conditional on the factor input matrix, it is essential in the present aggregation analysis. Using univariate statistics which depend on skewness and excess kurtosis, we find in most cases non-rejection of normality of log-inputs at the five per cent level; see Biørn *et al.* (2003b, Appendix D). However, for Pulp and paper, there is some evidence of rejection, especially at the start of the sample period. This is most pronounced for energy and materials, where normality is rejected at the 1 per cent level in the years 1972–1976. Despite these irregularities, we proceed by imposing normality of all log-inputs as a simplifying assumption in the application presented below.

9.5. *Empirical results*

9.5.1. *Estimates of exact-formulae moments, approximations and their components*

Utilizing (9.11) we have estimated the logs of expected output and of expected squared output for the 22 years in the sample period. The maximum, the mean and the minimum value of these annual time series are reported in the bottom row of Tables 9.3 and 9.4.[6] The formulae involve the mean vector and covariance matrix of the random parameters and the disturbance variance, estimated from the full panel data set, and the mean vector and covariance matrix of the log-inputs, calculated for each year. The rest of the two tables presents decompositions based on the approximation formula (9.14) transformed into logs, i.e., $\ln[G_1(Y)]$ and $\ln[G_2(Y)]$. For each of them, a total of seven components and their weights are specified. Again, due to space limitations, only the minimum, mean and maximum values obtained from the annual time series are reported.

Let us first consider *expected output*. The row labeled μ_y in Table 9.3 can be interpreted as mimicking the naïve way of representing the expectation of a log-normal variable, say W, as $\exp(\mathsf{E}[\ln(W)])$, and hence neglecting Jensen's inequality. This yields for Pulp and paper and Basic metals mean estimates of 4.124 and 3.629, which are considerably lower than the values obtained from the exact formulae, 6.230 and 6.938, respectively. The same is true for the other statistics. Including the 'variance

[6] Recall that the mean, unlike the minimum and maximum, is a linear operator.

Table 9.3. First-order moment estimates. Different approximations with components and exact-formula values. Summary statistics based on annual results, 1972–1993

		Pulp and paper			Basic metals		
	Components	Minimum	Mean	Maximum	Minimum	Mean	Maximum
μ_y	$\mu'_x\mu_\beta$	3.620	4.124	4.695	3.037	3.629	4.606
	$\frac{1}{2}\mu'_x\Sigma_{\beta\beta}\mu_x$	1.154	1.617	1.966	1.992	2.742	3.227
	$\frac{1}{2}\mu'_\beta\Sigma_{xx}\mu_\beta$	0.187	0.201	0.218	0.238	0.271	0.395
	$\frac{1}{2}\,\mathrm{tr}(\Sigma_{\beta\beta}\Sigma_{xx})$	0.102	0.124	0.140	0.123	0.161	0.329
	$\frac{1}{2}\sigma_{uu}$	0.020	0.020	0.020	0.049	0.049	0.049
$\ln[\Phi_1(y)]$[a]		5.804	6.087	6.334	6.379	6.852	7.224
	$\ln(\Gamma_1)$	−0.162	−0.104	−0.040	−0.330	−0.127	−0.071
	$\ln(\Lambda_1)$	0.111	0.217	0.309	0.094	0.186	0.280
$\ln[G_1(Y)]$[b]		5.899	6.201	6.469	6.426	6.911	7.304
$\ln[\mathsf{E}(Y)]$		5.927	6.230	6.500	6.440	6.938	7.333

[a] $\ln[\Phi_1(y)] = \mu'_x\mu_\beta + \frac{1}{2}(\mu'_x\Sigma_{\beta\beta}\mu_x + \mu'_\beta\Sigma_{xx}\mu_\beta + \mathrm{tr}(\Sigma_{\beta\beta}\Sigma_{xx}) + \sigma_{uu})$, cf. (9.7) and (9.16).

[b] $\ln[G_1(Y)] = \ln[\Phi_1(y)] + \ln(\Gamma_1) + \ln(\Lambda_1)$, cf. (9.15).

Table 9.4. Second-order moment estimates. Different approximations with components and exact-formula values. Summary statistics based on annual results, 1972–1993

		Pulp and paper			Basic metals		
	Components	Minimum	Mean	Maximum	Minimum	Mean	Maximum
$2\mu_y$	$2\mu'_x\mu_\beta$	7.239	8.249	9.390	6.074	7.257	9.212
	$2\mu'_x\Sigma_{\beta\beta}\mu_x$	4.617	6.469	7.862	7.969	10.968	12.908
	$2\mu'_\beta\Sigma_{xx}\mu_\beta$	0.748	0.806	0.871	0.952	1.082	1.578
	$2\,\mathrm{tr}(\Sigma_{\beta\beta}\Sigma_{xx})$	0.409	0.496	0.560	0.494	0.644	1.314
	$2\sigma_{uu}$	0.082	0.082	0.082	0.197	0.197	0.197
$\ln[\Phi_2(y)]$[a]		15.293	16.101	16.753	19.067	20.149	20.927
	$\ln(\Gamma_2)$	−1.297	−0.828	−0.318	−1.235	−0.937	−0.569
	$\ln(\Lambda_2)$	1.782	3.477	4.944	1.506	2.903	3.905
$\ln[G_2(Y)]$[b]		16.757	18.750	20.478	20.004	22.128	23.868
$\ln[\mathsf{E}(Y^2)]$		17.780	22.236	27.306	20.496	23.417	25.839

[a] $\ln[\Phi_2(y)] = 2(\mu'_x\mu_\beta + \mu'_x\Sigma_{\beta\beta}\mu_x + \mu'_\beta\Sigma_{xx}\mu_\beta + \mathrm{tr}(\Sigma_{\beta\beta}\Sigma_{xx}) + \sigma_{uu})$, cf. (9.7) and (9.16).

[b] $\ln[G_2(Y)] = \ln[\Phi_2(y)] + \ln(\Gamma_2) + \ln(\Lambda_2)$, cf. (9.15).

adjustment' which is part of the formula for the expectation of a log-normal variable, by considering $\ln[\Phi_1(y)]$, the means of the estimated expectations increase to 6.087 and 6.852, respectively, which are much closer to the values obtained from the exact formula. The same is true for the minimum and maximum values. A further decomposition of the contribution from the 'variance adjustment', into four components, is given in rows 2–5. They represent, respectively, coefficient heterogeneity, input variation, covariation between the two latter, and the genuine error term. It is worth noticing that for both industries the largest contribution (more than 80 per cent) comes from the coefficient variation, followed by input variation, interaction effects, whereas the smallest contribution (less than 2 per cent) comes from the variation in the error term.

Since output is not log-normally distributed marginally, there is a potential for a further improvement of the approximation, by also including the logs of the correction factors, $\ln[\Gamma_1]$ and $\ln[\Lambda_1]$. Summary statistics related to these factors are reported in rows 7 and 8 of Table 9.3. Whereas the mean of the estimates of the former is negative for both Pulp and paper and Basic metals, the mean of the latter is positive, as is also the case for their net effect. With one exception (the net effect for Basic metals in 1993) the above sign conclusions in fact hold for all years in both industries. The final results after including the contributions from these factors as well are reported in the row labeled $\ln[G_1(Y)]$.

An interesting question is how general these results are. Will, for instance, the qualitative conclusions carry over to other data sets? Since $\ln[\Lambda_1]$ is a positive definite quadratic form as long as $\Sigma_{\beta\beta}$ is a positive definite matrix, the contribution from this term will always be positive. If all the expected firm-level elasticities are positive and $\Sigma_{\beta\beta}$ and Σ_{xx} have only positive elements, also $\ln[\Gamma_1]$ would have given a positive contribution. However, when the signs of the off-diagonal elements of $\Sigma_{\beta\beta}$ and Σ_{xx} differ, negative estimates of $\ln[\Gamma_1]$ may occur.

As can be seen by comparing the last two rows in Table 9.3, the deviation from the mean based on the correct formula is modest in both industries, pointing to the fact that the approximation formula performs rather well in this case. Corresponding results hold for the maximum and minimum value. The approximate formula for log of expected output will be the point of departure in Section 9.5.2 when we will estimate two different measures of industry-level elasticities and compare the results to those relying on the representative agent construct.

We next turn to the corresponding approximation and decomposition results for *expected squared output*. This sheds light on the relative importance of different components with respect to *output volatility*. From the last two rows of Table 9.4 it is evident that the results based on the exact and on the approximate formulae differ more strongly for the second-order moments than for the first-order moments. Otherwise, the results are in line with those found for the first-order moments. Making the approximation gradually more sophisticated, by including more terms, starting with the naïve formula which disregards Jensen's inequality, we again get closer and closer to the exact-formula value of the moment. The row representing $\ln[\Phi_2(y)]$ in Table 9.4 mimics results obtained from the invalid assumption that output is log-normally distributed marginally. Including the logs of the correction factors, $\ln[\Gamma_2]$ and $\ln[\Lambda_2]$, brings us closer to the results obtained from the exact formula. The approximation seems to perform somewhat better for Basic metals than for Pulp and paper.

9.5.2. Aggregation biases in scale and input elasticities

Table 9.5 reports summary statistics for the industry-level elasticities obtained by the two hypothetic changes represented by (9.26). The underlying year specific elasticities are given in Table 9.6. In both industries and for all years, the estimated expected firm-level *scale elasticity* is smaller than the dispersion preserving scale elasticity, and larger than the mean preserving scale elasticity, but the discrepancies, i.e., the aggregation biases when sticking to naïve aggregation, are not very large. In Pulp and paper, the mean of the estimated dispersion preserving scale elasticity is 1.16, the estimated expected firm-level elasticity is 1.07 and the estimated mean preserving elasticity is 0.90. The maximal relative aggregation biases are about 9 and 25 per cent, when measured against the dispersion preserving and mean preserving elasticities, respectively. The corresponding elasticity estimates for Basic metals are 1.15, 1.11 and 1.00, respectively. The associated maximal relative biases are about 7 and 15 per cent. The annual variation in the industry elasticities is rather small.

However, the components of the scale elasticities, i.e., the *input elasticities*, show larger variability of the industry-level elasticities over the sample years. In some cases dramatic aggregation biases are found, relatively speaking. Irrespective of the definition and for both industries, the *materials elasticity* is the largest among the input elasticities. In all years, we find that not only for the scale elasticity, but also for materials, the estimate of the dispersion preserving elasticity exceeds the expected firm-level elasticity, which again exceeds the mean preserving elasticity. The maximal biases, relative to the dispersion preserving elasticity are about 21 and 7 per cent for Pulp and paper and Basic metals, respectively. When comparing with the mean preserving elasticity the corresponding relative biases are about 28 and 47 per cent. Overall, the two sets of estimated industry-level input elasticities show substantial variation over the sample period.

Table 9.5. Expected firm-level elasticities and absolute values of relative biases in per cent. Summary statistics based on annual results, 1972–1993

	Expected firm-level elasticity	Dispersion preserving ind.-level elasticity, bias in per cent			Mean preserving ind.-level elasticity, bias in per cent		
		Minimum	Mean	Maximum	Minimum	Mean	Maximum
Pulp and paper							
Scale	1.074	4.703	7.383	9.367	15.733	19.824	24.594
Capital	0.250	26.414	41.263	65.761	28.359	33.908	39.832
Labor	0.172	1.322	19.131	49.304	7.189	10.210	12.843
Energy	0.085	28.235	30.820	33.798	144.000	148.260	158.788
Materials	0.567	15.306	19.251	21.306	9.382	17.263	27.613
Basic metals							
Scale	1.106	0.353	4.073	6.816	8.229	10.525	14.860
Capital	0.125	1.111	129.207	1457.500	14.312	24.751	29.605
Labor	0.274	0.036	27.147	89.586	0.399	4.946	13.128
Energy	0.214	17.176	21.428	27.423	4.804	15.410	22.874
Materials	0.493	0.351	4.073	6.816	21.679	29.474	46.667

Turning to the relative magnitude of the specific input elasticity estimates, we find that whereas the firm-level *capital elasticity* is somewhat higher than the firm-level *labor elasticity* in Pulp and paper, and the dispersion preserving and the mean preserving aggregation experiment give approximately the same result. In contrast, in Basic metals the labor elasticity is substantially larger than the capital elasticity and the estimates of the industry-level elasticities show more variability. While in Pulp and paper the maximal relative biases are about 49 and 13 per cent for labor and about 66 and 40 per cent for capital, when related to the dispersion preserving and the mean preserving elasticity, respectively, the corresponding measures of the relative bias for Basic metals are 90 and 13 per cent for labor and still higher for capital. The dispersion preserving capital elasticity in the latter industry is very low at the start of the sample period, increases to about 0.15 in 1992 and again decreases substantially in the ultimate year. The mean bias exceeds 125 per cent. For the mean preserving elasticity we find a maximal bias of about 30 per cent.

The *energy elasticity* has, for Pulp and paper, the lowest estimate among the expected firm-level elasticities, 0.09. The dispersion preserving and mean preserving elasticities are about 0.12 and 0.03, respectively, and show almost no year-to-year variation. This corresponds to relative biases of about 34 and 159 per cent, respectively. In Basic metals the two aggregation experiments for the industry-level elasticities yield rather equal

Table 9.6. Industry-level scale and input elasticities, by year. Dispersion preserving and mean preserving values

	Dispersion preserving elasticity					Mean preserving elasticity				
Year	Scale	Capital	Labor	Energy	Materials	Scale	Capital	Labor	Energy	Materials
Pulp and paper										
1972	1.168	0.160	0.193	0.120	0.695	0.900	0.192	0.190	0.034	0.483
1973	1.166	0.156	0.186	0.125	0.699	0.896	0.192	0.188	0.035	0.481
1974	1.185	0.151	0.194	0.126	0.715	0.923	0.187	0.192	0.035	0.508
1975	1.181	0.163	0.191	0.125	0.703	0.925	0.183	0.191	0.035	0.515
1976	1.179	0.170	0.182	0.124	0.704	0.928	0.184	0.191	0.035	0.518
1977	1.172	0.192	0.192	0.119	0.669	0.901	0.192	0.195	0.034	0.480
1978	1.166	0.184	0.181	0.122	0.679	0.903	0.188	0.197	0.034	0.484
1979	1.165	0.173	0.177	0.125	0.690	0.899	0.187	0.195	0.035	0.482
1980	1.163	0.180	0.174	0.121	0.688	0.891	0.189	0.197	0.034	0.471
1981	1.162	0.183	0.160	0.122	0.696	0.895	0.189	0.194	0.034	0.477
1982	1.162	0.172	0.152	0.124	0.714	0.910	0.179	0.192	0.035	0.504
1983	1.155	0.186	0.148	0.120	0.701	0.889	0.188	0.191	0.034	0.477
1984	1.152	0.180	0.146	0.121	0.705	0.887	0.188	0.190	0.034	0.475
1985	1.154	0.178	0.141	0.123	0.713	0.887	0.189	0.191	0.034	0.474
1986	1.159	0.184	0.149	0.123	0.702	0.886	0.190	0.194	0.034	0.469
1987	1.159	0.183	0.144	0.123	0.709	0.896	0.185	0.192	0.034	0.486
1988	1.154	0.180	0.127	0.128	0.720	0.902	0.181	0.187	0.035	0.499
1989	1.151	0.181	0.127	0.127	0.716	0.893	0.183	0.188	0.035	0.486
1990	1.151	0.183	0.127	0.126	0.714	0.892	0.184	0.190	0.035	0.484
1991	1.150	0.185	0.120	0.129	0.716	0.895	0.183	0.189	0.035	0.487
1992	1.127	0.194	0.115	0.122	0.697	0.865	0.186	0.185	0.034	0.459
1993	1.134	0.198	0.117	0.122	0.697	0.862	0.195	0.189	0.033	0.444

(*Continued on next page*)

results. Estimation by the naïve approach produces maximal biases at about 27 and 23 per cent, relative to the dispersion preserving and the mean preserving elasticities, respectively.

Since neither the mean nor the covariance matrix of the log-input vector very rarely is time invariant, the assumptions underlying the dispersion and mean preserving aggregation experiments may seem too simplistic. It may be worthwhile to consider intermediate cases in which a weighting of the two extremes is involved. Some experiments along these lines suggest, contrary to what might be anticipated, that the expected firm-level elasticity is not uniformly closer to this weighted average than to either of the limiting cases. Still, the overall evidence from the above results, confined to the two synthetic aggregation experiments, gives a definite warning against using 'raw' firm-level elasticities to represent industry-level elasticities. As a basis for comparing patterns of productivity growth across countries, both naïvely aggregated input elasticities

Table 9.6. (Continued)

	Dispersion preserving elasticity					Mean preserving elasticity				
Year	Scale	Capital	Labor	Energy	Materials	Scale	Capital	Labor	Energy	Materials
Basic metals										
1972	1.180	0.008	0.360	0.206	0.606	1.002	0.167	0.270	0.186	0.379
1973	1.187	0.028	0.358	0.183	0.618	1.005	0.169	0.280	0.182	0.375
1974	1.177	0.056	0.325	0.169	0.627	0.994	0.177	0.260	0.174	0.383
1975	1.183	0.044	0.326	0.172	0.641	1.022	0.164	0.268	0.184	0.405
1976	1.174	0.049	0.326	0.189	0.610	1.004	0.161	0.278	0.195	0.371
1977	1.174	0.049	0.327	0.199	0.598	1.019	0.161	0.285	0.187	0.386
1978	1.152	0.096	0.266	0.187	0.604	1.002	0.173	0.254	0.178	0.396
1979	1.159	0.072	0.275	0.195	0.618	1.011	0.167	0.262	0.183	0.400
1980	1.162	0.073	0.290	0.196	0.603	0.998	0.165	0.276	0.191	0.366
1981	1.160	0.106	0.276	0.183	0.595	0.992	0.172	0.267	0.183	0.370
1982	1.151	0.140	0.240	0.165	0.606	0.991	0.175	0.258	0.180	0.379
1983	1.144	0.126	0.224	0.176	0.618	1.004	0.165	0.260	0.187	0.392
1984	1.148	0.116	0.237	0.183	0.613	0.995	0.165	0.268	0.191	0.371
1985	1.149	0.153	0.219	0.144	0.632	0.992	0.172	0.262	0.179	0.378
1986	1.144	0.143	0.213	0.161	0.627	0.988	0.169	0.259	0.186	0.374
1987	1.149	0.159	0.202	0.139	0.649	1.000	0.171	0.261	0.178	0.390
1988	1.145	0.150	0.190	0.139	0.667	1.004	0.170	0.253	0.179	0.403
1989	1.141	0.139	0.185	0.138	0.679	1.013	0.160	0.266	0.186	0.401
1990	1.138	0.144	0.175	0.147	0.672	1.006	0.163	0.257	0.187	0.398
1991	1.135	0.123	0.188	0.170	0.654	1.007	0.155	0.269	0.198	0.384
1992	1.113	0.154	0.145	0.174	0.640	0.963	0.163	0.243	0.204	0.353
1993	1.110	0.062	0.185	0.223	0.641	1.008	0.109	0.306	0.257	0.336

and time-invariant elasticities estimated solely from aggregate time series of output and inputs are potentially misleading.

9.6. Conclusion and extensions

This paper has been concerned with the aggregation of micro Cobb–Douglas production functions to the industry level when the firm specific production function parameters and the log-inputs are assumed to be independent and multinormally distributed. First, we have provided analytical approximations for the expectation and the higher-order origo moments of output, as well as conditions for the existence of such moments. These existence conditions turn out to be rather strong in the present case: only the first- and second-order moments exist. To some extent, this is due to our simplifying normality assumption, so that products of two vectors, both with support extending from minus to plus infinity, will enter the exponent of the expression for the moments of output. This suggests that investigating truncated distributions, in particular for the coefficients, may be an interesting topic for further research. Relaxation of normality and/or truncation is, however, likely to increase the analytical and numerical complexity of the aggregation procedures.

Second, we have shown how an industry-level production function, expressed as a relationship between expected output and expected inputs, can be derived and how discrepancies between correctly aggregated input and scale elasticities and their expected counterparts obtained from micro data can be quantified. It is quite obvious that the non-linearity of the mean production function combined with random coefficient variation across firms implies that the correctly aggregated coefficients in the 'aggregate Cobb–Douglas production function' are not strict technology parameters – not even to an acceptable degree of approximation – as they also depend on the coefficient heterogeneity and the covariance matrix of the log-input vector; cf. the quotation from Felipe and Fisher (2003, p. 209) in the introduction. The parameters characterizing our non-linear micro structure are not recoverable from time series of linearly aggregated output and input volumes. If agencies producing aggregate data could furnish macro-economists not only with simple sums and arithmetic means, but also with time series for other aggregates, say means, variances and covariances of logged variables, coefficients of variation, etc., they would have the opportunity to go further along the lines we have indicated. Our empirical decompositions have given evidence of this. Anyway, our results may provide guidance in situations where only aggregated data are available, and where applications such as forecasting of productivity changes and policy analysis could benefit from undertaking sensitivity analysis.

To indicate the possible range of the appropriately aggregated Cobb–Douglas parameters, we have provided results for the limiting dispersion preserving and mean preserving cases. However, the experiment underlying our definition of the mean preserving elasticities is one in which the variances of the log-inputs, but none of their covariances, are allowed to change. This simplifying assumption may have affected some of the above conclusions. An interesting alternative may be to assume that the correlation matrix of the log-input vector, rather than the covariances, is invariant when the variances change.

The dispersion preserving scale elasticity is substantially higher than the expected firm-level scale elasticity for both industries and in all the years. For the mean preserving counterpart the differences are smaller: for Pulp and paper the firm-level elasticity exceeds the aggregate elasticity in all years. It is worth noting that the ranking of the industry-level and the expected firm-level input elasticities do not coincide, and in addition, the former changes over time.

An assumption not put into question in this paper is zero correlation between the production function parameters and the log-inputs. An interesting extension would be to relax this assumption, for instance to model the correlation. Simply treating all parameters as fixed and firm specific would, however, imply wasting a substantial part of the sample, since a minimal time series length is needed to estimate firm specific fixed parameters properly.

Whether an extension of our approach to more flexible micro technologies, like the CES, the Translog, or the Generalized Leontief production functions, is practicable is an open question. First, exact-moment formulae will often not exist in closed form, and it may be harder both to obtain useful analytical approximations for expected output and to verify and ensure the existence of relevant moments. Second, if the two normality assumptions are retained, the problems of non-existence of higher-order moments are likely to increase since, for example, the Translog and the Generalized Leontief functions contain second-order terms. On the other hand, relaxing normality, in favor of truncated or other less heavy-tailed parametric distributions, will most likely increase the analytical complexity further. Then abandoning the full parametric approach may be the only way out.

Acknowledgements

Earlier versions of the paper were presented at the 2nd Nordic Econometric Meeting, Bergen, May 2003, the 11th International Panel Data Conference, College Station (TX), June 2004, and the 19th Annual Congress of the European Economic Association, Madrid, August 2004. We thank the referee for detailed and thoughtful comments, and Johan Heldal, Zhiyang Jia, and Dag Einar Sommervoll for helpful and clarifying remarks.

Appendix A9. Proofs

A9.1. Proof of Equation (9.11)

Inserting for the density of δ,

$$f(\delta) = (2\pi)^{-(n+1)/2}|\Sigma_{\beta\beta}|^{-1/2}\exp\left[-\frac{1}{2}\delta'\Sigma_{\beta\beta}^{-1}\delta\right],$$

we find that the last expectation in (9.10) can be written as

$$H_r = \mathsf{E}_\delta\left\{\exp\left[\left(r\mu_x' + r^2\mu_\beta'\Sigma_{xx}\right)\delta + \frac{1}{2}r^2\delta'\Sigma_{xx}\delta\right]\right\}$$

$$= \int_{R^{n+1}} \exp\left[a(r)'\delta + \frac{1}{2}r^2\delta'\Sigma_{xx}\delta\right] f(\delta)\,\mathrm{d}\delta$$
$$= (2\pi)^{-(n+1)/2}|\Sigma_{\beta\beta}|^{-1/2} \int_{R^{n+1}} \exp\left[\lambda_r(\delta)\right]\mathrm{d}\delta, \tag{A9.1}$$

where $a(r) = r\mu_x + r^2\Sigma_{xx}\mu_\beta$, $M(r) = \Sigma_{\beta\beta}^{-1} - r^2\Sigma_{xx}$, and

$$\lambda_r(\delta) = a(r)'\delta - \frac{1}{2}\delta' M(r)\delta = \frac{1}{2}a(r)'M(r)^{-1}a(r)$$
$$- \frac{1}{2}\left[\delta' - a(r)'M(r)^{-1}\right]M(r)\left[\delta - M(r)^{-1}a(r)\right]. \tag{A9.2}$$

Since integration goes over R^{n+1}, we can substitute $q = \delta - M(r)^{-1}a(r)$, giving

$$H_r = |\Sigma_{\beta\beta}|^{-1/2}\exp\left[\frac{1}{2}a(r)'M(r)^{-1}a(r)\right]$$
$$\times \int_{R^{n+1}} (2\pi)^{-(n+1)/2}\exp\left[-1/2q'M(r)q\right]\mathrm{d}q.$$

The integrand resembles a normal density function, with $M(r)$ occupying the same place as the *inverse* of the covariance matrix of q. Thus, the latter integral after division by $|M(r)^{-1}|^{1/2}$ equals one for any q and any positive definite $M(r)$, which implies

$$\int_{R^{n+1}} (2\pi)^{-(n+1)/2}\exp\left[-\frac{1}{2}q'M(r)q\right]\mathrm{d}q = \left|M(r)^{-1}\right|^{1/2}.$$

We can then express H_r in closed form as

$$H_r = |\Sigma_{\beta\beta}|^{-1/2}\exp\left[\frac{1}{2}a(r)'M(r)^{-1}a(r)\right]\left|M(r)\right|^{-1/2}$$
$$= \exp\left[\frac{1}{2}a(r)'M(r)^{-1}a(r)\right]\left|\Sigma_{\beta\beta}M(r)\right|^{-1/2},$$

which, inserted into (A9.1), yields

$$\mathsf{E}\left(Y^r\right) = \left|M(r)\Sigma_{\beta\beta}\right|^{-1/2}$$
$$\times \exp\left[r\mu_x'\mu_\beta + \frac{1}{2}r^2(\mu_\beta'\Sigma_{xx}\mu_\beta + \sigma_{uu}) + \frac{1}{2}a(r)'M(r)^{-1}a(r)\right]. \tag{A9.3}$$

Inserting for $a(r)$ and $M(r)$ completes the proof.

A9.2. Proof of Equation (9.26)

The first three components of $\ln[G_1(Y)]$, as given by (9.25), respond to changes in μ_z and the last three elements respond to changes in Σ_{zz}. Inserting in (9.7) from (9.2) and (9.3), we obtain

$$\begin{aligned}
&\sigma_{yy} = \sigma_{\alpha\alpha} + 2\mu_z\sigma_{\gamma\alpha} + \mu_z'\Sigma_{\gamma\gamma}\mu_z + \mu_\gamma'\Sigma_{zz}\mu_\gamma + \operatorname{tr}(\Sigma_{\gamma\gamma}\Sigma_{zz}) + \sigma_{uu},\\
&\mu_x'\Sigma_{\beta\beta}\Sigma_{xx}\mu_\beta = \sigma_{\gamma\alpha}'\Sigma_{zz}\mu_\gamma + \mu_z'\Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma,\\
&\mu_\beta'\Sigma_{xx}\Sigma_{\beta\beta}\Sigma_{xx}\mu_\beta = \mu_\gamma'\Sigma_{zz}\Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma.
\end{aligned}$$

Differentiating the various terms in (9.25) with respect to μ_z and Σ_{zz} (Lütkepohl, 1996, Section 10.3.2, Equations (2), (5) and (21)) we get

$$\frac{\partial\mu_y}{\partial\mu_z} = \frac{\partial(\mu_x'\mu_\beta)}{\partial\mu_z} = \frac{\partial(\mu_z'\mu_\gamma)}{\partial\mu_z} = \mu_\gamma, \tag{A9.4}$$

$$\frac{\partial\sigma_{yy}}{\partial\mu_z} = 2\frac{\partial(\mu_z'\sigma_{\alpha\gamma})}{\partial\mu_z} + \frac{\partial(\mu_z'\Sigma_{\gamma\gamma}\mu_z)}{\partial\mu_z} = 2(\sigma_{\gamma\alpha} + \Sigma_{\gamma\gamma}\mu_z), \tag{A9.5}$$

$$\frac{\partial(\mu_x'\Sigma_{\beta\beta}\Sigma_{xx}\mu_\beta)}{\partial\mu_z} = \frac{\partial(\mu_z'\Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma)}{\partial\mu_z} = \Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma, \tag{A9.6}$$

$$\frac{\partial\sigma_{yy}}{\partial\Sigma_{zz}} = \frac{\partial(\mu_\gamma'\Sigma_{zz}\mu_\gamma)}{\partial\Sigma_{zz}} + \frac{\partial\operatorname{tr}(\Sigma_{\gamma\gamma}\Sigma_{zz})}{\partial\Sigma_{zz}} = \mu_\gamma\mu_\gamma' + \Sigma_{\gamma\gamma}, \tag{A9.7}$$

$$\begin{aligned}
\frac{\partial(\mu_x'\Sigma_{\beta\beta}\Sigma_{xx}\mu_\beta)}{\partial\Sigma_{zz}} &= \frac{\partial(\sigma_{\gamma\alpha}'\Sigma_{zz}\mu_\gamma)}{\partial\Sigma_{zz}} + \frac{\partial(\mu_z'\Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma)}{\partial\Sigma_{zz}}\\
&= \frac{\partial\operatorname{tr}(\sigma_{\gamma\alpha}'\Sigma_{zz}\mu_\gamma)}{\partial\Sigma_{zz}} + \frac{\partial\operatorname{tr}(\mu_z'\Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma)}{\partial\Sigma_{zz}} = \mu_\gamma\sigma_{\gamma\alpha}' + \mu_\gamma\mu_z'\Sigma_{\gamma\gamma},
\end{aligned} \tag{A9.8}$$

$$\begin{aligned}
\frac{\partial(\mu_\beta'\Sigma_{xx}\Sigma_{\beta\beta}\Sigma_{xx}\mu_\beta)}{\partial\Sigma_{zz}} &= \frac{\partial(\mu_\gamma'\Sigma_{zz}\Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma)}{\partial\Sigma_{zz}}\\
&= \frac{\partial\operatorname{tr}(\mu_\gamma'\Sigma_{zz}\Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma)}{\partial\Sigma_{zz}} = \mu_\gamma\mu_\gamma'\Sigma_{zz}\Sigma_{\gamma\gamma} + \Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma\mu_\gamma'.
\end{aligned} \tag{A9.9}$$

It follows from (9.25) and (A9.4)–(A9.9), that

$$\frac{\partial\ln[G_1(Y)]}{\partial\mu_z} = \mu_\gamma + \sigma_{\gamma\alpha} + \Sigma_{\gamma\gamma}(\mu_z + \Sigma_{zz}\mu_\gamma), \tag{A9.10}$$

$$\begin{aligned}
\frac{\partial\ln[G_1(Y)]}{\partial\Sigma_{zz}} &= \frac{1}{2}(\mu_\gamma\mu_\gamma' + \Sigma_{\gamma\gamma} + \mu_\gamma\mu_\gamma'\Sigma_{zz}\Sigma_{\gamma\gamma}\\
&\quad + \Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma\mu_\gamma') + \mu_\gamma(\sigma_{\gamma\alpha}' + \mu_z'\Sigma_{\gamma\gamma}).
\end{aligned} \tag{A9.11}$$

Since, from (9.19), $\Delta \ln[\mathsf{E}(Z)] = \Delta(\mu_z + \frac{1}{2}\sigma_{zz})$, we have

$$\frac{\partial \ln[G_1(Y)]}{\partial \ln[\mathsf{E}(Z)]} = \begin{cases} \mu_\gamma + \sigma_{\gamma\alpha} + \Sigma_{\gamma\gamma}(\mu_z + \Sigma_{zz}\mu_\gamma), & \text{when } \Sigma_{zz} \text{ is constant,} \\ \operatorname{diagv}[\mu_\gamma\mu_\gamma' + \Sigma_{\gamma\gamma} + \mu_\gamma\mu_\gamma'\Sigma_{zz}\Sigma_{\gamma\gamma} & \text{when } \mu_z \text{ and the} \\ \qquad + \Sigma_{\gamma\gamma}\Sigma_{zz}\mu_\gamma\mu_\gamma' & \quad\text{off-diagonal elements} \\ \qquad + 2\mu_\gamma(\sigma_{\gamma\alpha}' + \mu_z'\Sigma_{\gamma\gamma})], & \quad\text{of } \Sigma_{zz} \text{ are constant,} \end{cases} \tag{A9.12}$$

which completes the proof.

Appendix B9. Details on estimation and data

B9.1. Details on the ML estimation

We consider, for convenience, our unbalanced panel data set (cf. Appendix B9.2 below) as a data set where the firms are observed in at least 1 and at most P years, and arrange the observations in groups according to the time series lengths (a similar ordering is used in Biørn, 2004). Let N_p be the number of firms which are observed in p years (not necessarily the same and consecutive), let (ip) index the ith firm among those observed in p years, and let from now on t index the observation number $(t = 1, \ldots, p)$ rather than calendar time. The production function (9.1), can then be written as

$$\begin{aligned} &y_{(ip)t} = x_{(ip)t}'\beta_{(ip)} + u_{(ip)t}, \\ &\quad p = 1, \ldots, P;\ i = 1, \ldots, N_p;\ t = 1, \ldots, p, \end{aligned} \tag{B9.1}$$

where $\beta_{(ip)}$ is the coefficient vector of firm (ip). Inserting $\beta_{(ip)} = \mu_\beta + \delta_{(ip)}$ we get

$$y_{(ip)t} = x_{(ip)t}'\mu_\beta + \psi_{(ip)t}, \qquad \psi_{(ip)t} = x_{(ip)t}'\delta_{(ip)} + u_{(ip)t}. \tag{B9.2}$$

Stacking the p realizations from firm (ip) in $y_{(ip)} = [y_{(ip)1}, \ldots, y_{(ip)p}]'$, $X_{(ip)} = [x_{(ip)1}, \ldots, x_{(ip)p}]$, $u_{(ip)} = [u_{(ip)1}, \ldots, u_{(ip)p}]'$, and $\psi_{(ip)} = [\psi_{(ip)1}, \ldots, \psi_{(ip)p}]'$, we can write (B9.2) as

$$y_{(ip)} = X_{(ip)}'\mu_\beta + \psi_{(ip)}, \qquad \psi_{(ip)} = X_{(ip)}'\delta_{(ip)} + u_{(ip)}, \tag{B9.3}$$

where, from (9.2)–(9.4),

$$\psi_{(ip)}|X_{(ip)} \sim \mathcal{N}(0, \Omega_{(ip)}), \qquad \Omega_{(ip)} = X_{(ip)}'\Omega X_{(ip)} + \sigma_{uu}I_p, \tag{B9.4}$$

and I_p is the p-dimensional identity matrix. The log-likelihood function is therefore

$$\mathcal{L} = -\frac{m}{2}\ln(2\pi) - \frac{1}{2}\sum_{p=1}^{P}\sum_{i=1}^{N_p}\big\{\ln|\Omega_{(ip)}| \\ + [y_{(ip)} - X'_{(ip)}\mu_\beta]'\Omega_{(ip)}^{-1}[y_{(ip)} - X'_{(ip)}\mu_\beta]\big\}, \tag{B9.5}$$

where $m = \sum_{p=1}^{P} pN_p$. The ML estimators of $(\mu_\beta, \sigma_{uu}, \Omega)$ follow by maximizing $\mathcal{L}$. The solution may be simplified by concentrating $\mathcal{L}$ over μ_β and maximizing the resulting function with respect to σ_{uu} and the unknown elements of Ω.

B9.2. Data

The data are from the years 1972–1993 and represent two Norwegian manufacturing industries, Pulp and paper and Basic metals. Table B9.1, classifying the observations by the number of years, and Table B9.2, sorting the firms by the calendar year in which they are observed, shows the unbalanced structure of the data set. There is a negative trend in the number of firms for both industries.

The primary data source is the Manufacturing Statistics database of Statistics Norway, classified under the Standard Industrial Classification (SIC)-codes 341 Manufacture of paper and paper products (Pulp and paper, for short) and 37 Manufacture of basic metals (Basic metals, for short). Both firms with contiguous and non-contiguous time series are included. Observations with missing values of output or inputs have been removed. This reduced the effective sample size by 6–8 per cent in the two industries.

In the description below, MS indicates firm-level data from the Manufacturing Statistics, NNA indicates data from the Norwegian National Accounts, which are identical for firms classified in the same National Account industry.

Y: Output, 100 tonnes (MS)

$K = KB + KM$: Total capital stock (buildings/structures plus machinery/transport equipment), 100 000 1991-NOK (MS, NNA)

L: Labor input, 100 man-hours (MS)

E: Energy input, 100 000 kWh, electricity plus fuels (MS)

$M = CM/QM$: Input of materials, 100 000 1991-NOK (MS, NNA)

CM: Total materials cost (MS)

QM: Price of materials, 1991 = 1 (NNA)

Table B9.1. Number of firms (N_p) by number of replications (p)

	Pulp and paper		Basic metals	
p	N_p	$N_p p$	N_p	$N_p p$
22	60	1320	44	968
21	9	189	2	42
20	5	100	4	80
19	3	57	5	95
18	1	18	2	36
17	4	68	5	85
16	6	96	5	80
15	4	60	4	60
14	3	42	5	70
13	4	52	3	39
12	7	84	10	120
11	10	110	7	77
10	12	120	6	60
09	10	90	5	45
08	7	56	2	16
07	15	105	13	91
06	11	66	4	24
05	14	70	5	25
04	9	36	6	24
03	18	54	3	9
02	5	10	6	12
01	20	20	20	20
Sum	237	2823	166	2078

Output The firms in the Manufacturing Statistics are in general multi-output firms and report output of a number of products measured in both NOK and primarily tonnes or kg. For each firm, an aggregate output measure in tonnes is calculated. Hence, rather than representing output in the two industries by deflated sales, which may contain measurement errors (see Klette and Griliches, 1996), and recalling that the products from the two industries are relatively homogeneous, our output measures are actual output in physical units, which are in several respects preferable.

Capital stock The calculations of capital stock data are based on the perpetual inventory method, assuming constant depreciation rates. We combine firm data on gross investment with fire insurance values for each of the two categories Buildings and structures and Machinery and transport equipment from the MS. The data on investment and fire insurance are

Table B9.2. Number of firms by calendar year

Year	Pulp and paper	Basic metals
1972	171	102
1973	171	105
1974	179	105
1975	175	110
1976	172	109
1977	158	111
1978	155	109
1979	146	102
1980	144	100
1981	137	100
1982	129	99
1983	111	95
1984	108	87
1985	106	89
1986	104	84
1987	102	87
1988	100	85
1989	97	83
1990	99	81
1991	95	81
1992	83	71
1993	81	83
Sum	2823	2078

deflated using industry specific price indices of investment goods from the NNA (1991 = 1). The depreciation rate is set to 0.02 for Buildings and structures and 0.04 for Machinery and transport equipment. For further documentation, see Biørn *et al.* (2000, Section 4; 2003a).

Other inputs From the MS we get the number of man-hours used, total electricity consumption in kWh, the consumption of a number of fuels in various denominations, and total material costs in NOK for each firm. The different fuels are transformed to the common denominator kWh by using estimated average energy content of each fuel, which enables us to calculate aggregate energy use in kWh for each firm.

References

Biørn, E. (2004), “Regression systems for unbalanced panel data: a stepwise maximum likelihood procedure”, *Journal of Econometrics*, Vol. 122, pp. 281–291.

Biørn, E., Skjerpen, T. (2004), "Aggregation and aggregation biases in production functions: a panel data analysis of translog models", *Research in Economics*, Vol. 58, pp. 31–57.

Biørn, E., Lindquist, K.-G., Skjerpen, T. (2000), "Micro data on capital inputs: attempts to reconcile stock and flow information", Discussion Paper No. 268, Statistics Norway.

Biørn, E., Lindquist, K.-G., Skjerpen, T. (2002), "Heterogeneity in returns to scale: a random coefficient analysis with unbalanced panel data", *Journal of Productivity Analysis*, Vol. 18, pp. 39–57.

Biørn, E., Lindquist, K.-G., Skjerpen, T. (2003a), "Random coefficients in unbalanced panels: an application on data from chemical plants", *Annales d'Économie et de Statistique*, Vol. 69, pp. 55–83.

Biørn, E., Skjerpen, T., Wangen, K.R. (2003b), "Parametric aggregation of random coefficient Cobb–Douglas production functions: evidence from manufacturing production functions", Discussion Paper No. 342, Statistics Norway.

Embrechts, P., Klüppelberg, C., Mikosch, T. (1997), *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.

Evans, M., Hastings, N., Peacock, B. (1993), *Statistical Distributions*, 2nd ed., Wiley, New York.

Felipe, J., Fisher, F.M. (2003), "Aggregation in production functions: what applied economists should know", *Metroeconomica*, Vol. 54, pp. 208–262.

Fortin, N.M. (1991), "Fonctions de production et biais d'agrégation", *Annales d'Économie et de Statistique*, Vol. 20/21, pp. 41–68.

Hildenbrand, W. (1998), "How relevant are specifications of behavioral relations on the micro level for modelling the time path of population aggregates?", *European Economic Review*, Vol. 42, pp. 437–458.

Jorgenson, D.W. (1995), *Productivity. Volume 2: International Comparisons of Economic Growth*, MIT Press, Cambridge.

Klette, T.J., Griliches, Z. (1996), "The inconsistency of common scale estimators when output prices are unobserved and endogenous", *Journal of Applied Econometrics*, Vol. 11, pp. 343–361.

Lewbel, A. (1990), "Income distribution movements and aggregate money illusion", *Journal of Econometrics*, Vol. 43, pp. 35–42.

Lewbel, A. (1992), "Aggregation with log-linear models", *Review of Economic Studies*, Vol. 59, pp. 635–642.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger R.D. (1996), *SAS System for Mixed Models*, SAS Institute, Cary.

Lütkepohl, H. (1996), *Handbook of Matrices*, Wiley, Chichester.

Mas-Colell, A., Whinston, M.D., Green, J.R. (1995), *Microeconomic Theory*, Oxford University Press, New York.

Stoker, T.M. (1986a), "Aggregation, efficiency, and cross-section regression", *Econometrica*, Vol. 54, pp. 171–188.

Stoker, T.M. (1986b), "Simple tests of distributional effects on macroeconomic equations", *Journal of Political Economy*, Vol. 94, pp. 763–795.

Stoker, T.M. (1993), "Empirical approaches to the problem of aggregation over individuals", *Journal of Economic Literature*, Vol. 31, pp. 1827–1874.

van Garderen, K.J., Lee, K., Pesaran, M.H. (2000), "Cross-sectional aggregation of nonlinear models", *Journal of Econometrics*, Vol. 95, pp. 285–331.

Zellner, A. (1969), "On the aggregation problem: a new approach to a troublesome problem", in: Fox, K.A., Sengupta, J.K., Narasimham, G.V.L., editors, *Economic Models, Estimation and Risk Programming: Essays in Honor of Gerhard Tintner*, Springer, Berlin, pp. 365–378.

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74010-X

CHAPTER 10

Conditional Heteroskedasticity and Cross-Sectional Dependence in Panel Data: An Empirical Study of Inflation Uncertainty in the G7 countries

Rodolfo Cermeño[a] and Kevin B. Grier[b]

[a]División de Economía, CIDE, México D.F., México
E-mail address: rodolfo.cermeno@cide.edu
[b]Department of Economics, University of Oklahoma, OK 73019, USA
E-mail address: angus@ou.edu

Abstract

Despite the significant growth of macroeconomic and financial empirical panel studies the modeling of time dependent variance–covariance processes has not yet been addressed in the panel data literature. In this paper we specify a model that accounts for conditional heteroskedasticity and cross-sectional dependence within a typical panel data framework. We apply the model to a panel of monthly inflation rates of the G7 countries over the period 1978.2–2003.9 and find significant and quite persistent patterns of volatility and cross-sectional dependence. We then use the model to test two hypotheses about the interrelationship between inflation and inflation uncertainty, finding no support for the hypothesis that higher inflation uncertainty produces higher average inflation rates and strong support for the hypothesis that higher inflation is less predictable.

Keywords: dynamic panel data models, conditional heteroskedasticity, cross-sectional dependence, GARCH models, inflation uncertainty

JEL classifications: C33, C15

10.1. Introduction

The empirical panel data literature on financial and macroeconomic issues has grown considerably in the few past years. A recent search of

ECONLIT using the keyword phrases “financial panel data” and “macroeconomic panel data” produced 687 and 309 hits respectively.[1] While it is well known that most financial and macroeconomic time series data are conditionally heteroskedastic, rendering traditional estimators consistent, but inefficient, this rapidly growing literature has not yet addressed the issue. On the other hand, sophisticated multivariate GARCH models already are in wide use but they are confined to a time series context.[2]

In this paper we specify a panel model that accounts for conditional heteroskedasticity and cross-sectional correlation. The model is used to characterize the patterns of volatility and cross-sectional dependence of inflation in the G7 countries and to evaluate the hypotheses that (i) higher inflation uncertainty produces higher average inflation rates and (ii) higher inflation rates become less predictable. The main contribution of the paper is to account for a time dependent error covariance processes in panel models with fixed effects (dynamic or static), thus opening an avenue for empirical panel research of financial or macroeconomic volatility.

Although the volatility processes can be studied on an individual basis (i.e. country by country) using existing GARCH models (e.g., Engle, 1982; Engle *et al.*, 1987; Bollerslev *et al.*, 1988; Bollerslev, 1990), panel modeling is still worth pursuing since taking into account the cross-sectional dependence will increase efficiency and provide potentially important information about patterns of cross-sectional dependence.

It is important to remark, though, that identification of time dependent variance–covariance processes in panel data is feasible as long as the cross-sectional dimension N is relatively small since the number of covariance parameters will increase rapidly otherwise, which limits the applicability of the model to relatively small N and large T panels.[3]

The rest of the paper is organized as follows. In Section 10.2 we formulate the basic panel model with conditional heteroskedastic and cross-sectionally correlated disturbances and briefly discuss some special cases and generalizations. Section 10.3 discusses the strategy that will be followed in order to determine the presence of time dependent variance–covariance processes and to specify a preliminary panel model with such effects. Section 10.4 provides the empirical results, characterizing volatility and cross-sectional dependence in the G7 countries, as well as testing two hypotheses about the interrelationship between inflation and its predictability. Finally, Section 10.5 concludes.

[1] Search conducted November 8, 2004.

[2] See Bollerslev *et al.* (1992) for a survey on ARCH models. For a comprehensive survey on multivariate GARCH models see Bauwens *et al.* (2003).

[3] Phillips and Sul (2003) point out this limitation in the context of heterogeneous panels with (unconditional) cross-sectional dependence.

10.2. The model

Consider the following dynamic panel data (DPD) model with fixed effects:[4]

$$y_{it} = \mu_i + \phi y_{it-1} + \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{10.1}$$

where N and T are the number of cross sections and time periods respectively; y_{it} is the dependent variable, μ_i is an individual specific effect, which is assumed fixed, $\mathbf{x}_{it}$ is a row vector of exogenous explanatory variables of dimension k, and $\boldsymbol{\beta}$ is a k by 1 vector of coefficients. We assume that the AR parameter satisfies the condition $|\phi| < 1$ and that T is relatively large so that we can invoke consistency of the Least Squares estimators.[5] In the case $\phi = 0$, the process given by Equation (10.1) becomes static.[6] The disturbance term u_{it} is assumed to have a zero mean normal distribution with the following conditional moments:

$$\begin{aligned}
&\text{(i) } E[u_{it}u_{js}/u_{it-1}, u_{js-1}] = \sigma^2_{it} \quad \text{for } i = j \text{ and } t = s,\\
&\text{(ii) } E[u_{it}u_{js}/u_{it-1}, u_{js-1}] = \sigma_{ijt} \quad \text{for } i \neq j \text{ and } t = s,\\
&\text{(iii) } E[u_{it}u_{js}/u_{it-1}, u_{js-1}] = 0 \quad \text{for } i = j \text{ and } t \neq s,\\
&\text{(iv) } E[u_{it}u_{js}/u_{it-1}, u_{js-1}] = 0 \quad \text{for } i \neq j \text{ and } t \neq s.
\end{aligned} \tag{10.2}$$

Assumption (iii) states that there is no autocorrelation while assumption (iv) disallows non-contemporaneous cross-sectional correlation.[7] Assumptions (i) and (ii) define a very general conditional variance–covariance process; some structure needs to be imposed in order to make this process tractable. We propose the following specification which is an adaptation of the model in Bollerslev *et al.* (1988).

$$\sigma^2_{it} = \alpha_i + \delta\sigma^2_{i,t-1} + \gamma u^2_{i,t-1}, \quad i = 1, \ldots, N, \tag{10.3}$$

$$\sigma_{ijt} = \eta_{ij} + \lambda\sigma_{ij,t-1} + \rho u_{i,t-1}u_{j,t-1}, \quad i \neq j. \tag{10.4}$$

The model defined by Equations (10.1) (conditional mean), (10.3) (conditional variance) and (10.4) (conditional covariance) is simply a DPD

[4] This class of models is widely known in the panel data literature. See Baltagi (2001) and Hsiao (2003) for details.

[5] For dynamic models with fixed effects and i.i.d. errors, it is well known that the LSDV estimator is downward biased in small T samples. See, for example, Kiviet (1995).

[6] It is worth emphasizing that we are only considering the case of stationary panels. In practice, we will have to assure that all variables are indeed stationary or $I(0)$.

[7] Ruling out autocorrelation might be a restrictive assumption but it is convenient because of its simplicity. In practice, we will need to make sure that this assumption is not violated.

model with conditional covariance. Thus, we can use the acronym DPD-CCV.[8] Modeling the conditional variance and covariance processes in this way is quite convenient in a panel data context since by imposing a common dynamics to each of them, the number of parameters is considerably reduced. In this case there are $(\frac{1}{2}N(N+1)+4)$ parameters in the covariance matrix. It is important to emphasize that (10.3) and (10.4) imply that the conditional variance and covariance processes follow, respectively, a common dynamics but their actual values, however, are not identical for each unit or pair of units (conditionally or unconditionally).

It can be shown that the conditions $\alpha_i > 0$, $(\delta+\gamma) < 1$, and $(\lambda+\rho) < 1$ are sufficient for the conditional variance and covariance processes to converge to some fixed (positive in the case of the variance) values. However, in general there is no guarantee that the covariance matrix of disturbances be positive definite (at each point in time) and that it converges to some fixed positive definite matrix. Thus, assuming positive definiteness of the covariance matrix, the error structure of the model will reduce, unconditionally, to the well-known case of groupwise heteroskedasticity and cross-sectional correlation.

In matrix notation and assuming given initial values y_{i0}, Equation (10.1) becomes

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{Z}_t\boldsymbol{\theta} + \mathbf{u}_t, \quad t = 1, \dots, T, \tag{10.5}$$

where $\mathbf{y}_t$, $\mathbf{u}_t$, are vectors of dimension $N \times 1$. The matrix $\mathbf{Z}_t = [\mathbf{y}_{t-1} \vdots \mathbf{X}_t]$ has dimension $N \times (K+1)$, $\boldsymbol{\mu}$ is a $N \times 1$ vector of individual specific effects, and $\boldsymbol{\theta} = [\phi \vdots \boldsymbol{\beta}']'$ is a conformable column vector of coefficients. Given our previous assumptions the N-dimensional vector of disturbances $\mathbf{u}_t$ will follow a zero-mean multivariate normal distribution, denoted as $\mathbf{u}_t \sim N(\mathbf{0}, \boldsymbol{\Omega}_t)$. The covariance matrix $\boldsymbol{\Omega}_t$ is time dependent and its diagonal and off-diagonal elements are given by Equations (10.3) and (10.4) respectively. The vector of observations $\mathbf{y}_t$ is therefore conditionally normally distributed with mean $(\boldsymbol{\mu} + \mathbf{Z}_t\boldsymbol{\theta})$ and variance–covariance matrix $\boldsymbol{\Omega}_t$. That is, $\mathbf{y}_t \sim N(\boldsymbol{\mu} + \mathbf{Z}_t\boldsymbol{\theta}, \boldsymbol{\Omega}_t)$ and its conditional density is

$$\begin{aligned} f(\mathbf{y}_t/\mathbf{Z}_t, \boldsymbol{\mu}, \boldsymbol{\theta}, \varphi) = (2\pi)^{-N/2}|\boldsymbol{\Omega}_t|^{-1/2} \exp\left(-\frac{1}{2}\right)(\mathbf{y}_t - \boldsymbol{\mu} - \mathbf{Z}_t\boldsymbol{\theta})' \\ \times \boldsymbol{\Omega}_t^{-1}(\mathbf{y}_t - \boldsymbol{\mu} - \mathbf{Z}_t\boldsymbol{\theta}), \end{aligned} \tag{10.6}$$

[8] We should remark that Equations (10.3) and (10.4) could have a more general GARCH (p, q) formulation.

where φ includes the parameters in Equations (10.3) and (10.4). For the complete panel we have the following log-likelihood function:[9]

$$l = -\left(\frac{NT}{2}\right)\ln(2\pi) - \left(\frac{1}{2}\right)\sum_{t=1}^{T}\ln|\boldsymbol{\Omega}_t| - \left(\frac{1}{2}\right)\sum_{t=1}^{T}(\mathbf{y}_t - \boldsymbol{\mu} - \mathbf{Z}_t\boldsymbol{\theta})' \times \boldsymbol{\Omega}_t^{-1}(\mathbf{y}_t - \boldsymbol{\mu} - \mathbf{Z}_t\boldsymbol{\theta}). \tag{10.7}$$

This function is similar to those derived in the context of multivariate GARCH models (e.g., Bollerslev *et al.*, 1988).[10] It can be shown straightforwardly that if the disturbances are cross-sectionally independent the $N \times N$ matrix $\boldsymbol{\Omega}_t$ becomes diagonal and the log-likelihood function takes the simpler form:

$$l = -\frac{NT}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\ln\left(\sigma_{it}^2(\varphi)\right) - \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{(y_{it} - \mu_i - \phi y_{it-1} - \mathbf{x}_{it}\beta)^2}{\sigma_{it}^2(\varphi)}. \tag{10.8}$$

Further, in the absence of conditional heteroskedasticity and cross-sectional correlation the model simply reduces to a typical DPD model.

Even though the LSDV estimator in Equation (10.1) is still consistent it will no longer be efficient in the presence of conditional heteroskedastic and cross-sectionally correlated errors, either conditionally or unconditionally. In this case, the proposed non-linear MLE estimator based upon (10.7) or (10.8) (depending on whether we have cross-sectionally correlated disturbances or not) will be appropriate. Note that, by using the MLE estimator we are able to obtain both the parameters of the conditional mean and conditional variance–covariance equations while the LSDV estimator will only be able to compute the coefficients in the mean equation.

It is well known that under regularity conditions the MLE estimator is consistent, asymptotically efficient and asymptotically normally distributed. Also it is known that these properties carry through when the

[9] It should be remarked that the normality assumption may not hold in practice leading to Quasi-MLE estimation. See Davidson and McKinnon (1993) for a general discussion. Although this issue needs further investigation it is worth pointing out that Bollerslev and Wooldridge (1992) find that the finite sample biases in the QMLE appear to be relatively small in time series GARCH models.

[10] Also it is similar to the log likelihood function derived in the context of prediction error decomposition models for multivariate time series. See for example Brockwell and Davis (1991) and Harvey (1990).

observations are time dependent as is the case of multivariate GARCH processes. Therefore, the MLE estimator in (10.7) or (10.8) is asymptotically normally distributed with mean equal to the true parameter vector and a covariance matrix equal to the inverse of the corresponding information matrix. It is important to note that these asymptotic properties would hold for N fixed and T approaching to infinity since we are modeling the N-dimensional vector of disturbances of the panel as a multivariate time series process.

Estimation of the DPDCCV model will be made by direct maximization of the log-likelihood function given by (10.7), using numerical methods.[11] The asymptotic covariance matrix of the MLE estimator of this type will be approximated by the negative inverse of the Hessian of l evaluated at MLE parameter estimates. It is important to remark that the total number of coefficients to be estimated depends on the squared cross-sectional dimension of the panel, N^2, which in practice suggests applying the model to relatively small N panels in order to make the estimation feasible and to retain the asymptotic properties, namely consistency and efficiency, of this estimator.[12]

In practice, the individual effects in the mean equation may not be significantly different from each other giving rise to a mean equation with a single intercept (often called “pooled regression model”). Also, it is possible that the conditional variance or covariance processes do not exhibit individual effects. A combination of these possibilities could occur as well. A completely heterogeneous panel with individual specific coefficients for all the parameters in the mean and variance–covariance equations can also be considered, although in this last case we can run into estimation problems given the considerably large number of parameters that will arise even if the number of cross sections is relatively small.

Finally, it is worth mentioning some alternative specifications for the variance and covariance processes along the lines of those developed in the multivariate GARCH literature. For example, a variation of Equation (10.4) that specifies the analogous of the constant correlation model as in Bollerslev (1990) or its generalized version, the dynamic conditional correlation model, given in Engle (2002). Also, depending on the particular subject of study, exogenous regressors can be included in the variance equations as well as the variance itself can be included as a regressor in the conditional mean equation, as in multivariate M-GARCH-M models.

[11] We use the GAUSS Optimization module.

[12] In this paper we only consider small values of N. Further work will focus on using existing multivariate GARCH two-step methods which allow consistent, although inefficient, estimation of a considerably large number of parameters as would be the case of larger N panels. See Engle (2002), and Ledoit *et al*. (2003).

10.3. Empirical strategy

Since the proposed DPDCCV models are non-linear and estimation by direct maximization of the log-likelihood can be tedious work, it may be helpful to make some preliminary identification of the most appropriate model. In what follows we outline an empirical methodology for this purpose although it should be remarked that it is only done in an informal way.

Two issues are fundamental in our empirical strategy: (i) Specifying the best model for the mean equation and (ii) Identifying conditional variance–covariance processes in the panel. We consider that, provided there are a large enough number of time series observations so that we can rely on consistency of LS estimators, these issues can be addressed using conventional panel data estimation results as we discuss next.

10.3.1. Specifying the mean equation

An important issue in empirical panel work is the poolability of the data. In the context of Equation (10.1) we need to determine whether there are individual specific effects or a single intercept.[13] For this purpose we can test for individual effects in the mean equation using the LSDV estimator with a heteroskedasticity and autocorrelation consistent covariance matrix, along the lines of White (1980) and Newey and West (1987) estimators applied to panel.[14]

Under the assumption of cross-sectional independence, and for models where the variance process is identical across units, the LSDV and OLS estimators respectively are still best linear estimators. However if the variances are not equal across units the unconditional variance process will differ across units and the previous estimators will no longer be efficient. Given that we do not know a priori the appropriate model and that we may have autocorrelation problems in practice, it seems convenient to use a covariance matrix robust to heteroskedasticity and autocorrelation. Specifically, we can test the null hypothesis $H_0\colon \mu_1 = \mu_2 = \cdots = \mu_N$ by means of a Wald-test, which will follow a $\chi^2_{(N-1)}$ distribution asymptotically.

[13] From a much broader perspective, however, we need to determine if full heterogeneity or some form of pooling is more appropriate for the conditional mean equation.

[14] Arellano (1987) has extended White's heteroskedasticity consistent covariance estimator to panel data but this estimator is not appropriate here since it has been formulated for small T and large N panels which is not our case.

10.3.2. Identifying conditional variance–covariance processes

Once we have determined a preliminary model for the mean equation, we can explore the possibility of a time dependent pattern in the variance process by examining whether the squared LSDV or LS residuals (depending on whether individual specific effects are included or not in the mean equation) exhibit a significant autocorrelation pattern.[15] Depending upon the number of significant partial autocorrelations obtained we can choose a preliminary order for the variance process. As a practical rule, we can consider an ARCH (1) process if only the first lag is significant, or a GARCH (1, 1) if more lags are significant.

A related important issue is to determine if there are individual effects in the variance process. This can be done by testing for individual effects in the AR regression of squared residuals. Complementarily, a test for unconditional groupwise heteroskedasticity (which can be done in a conventional way) can lead us to decide for individual effects in the variance process if the null hypothesis is rejected.

Next, we can carry out a conventional test for the null hypothesis of no cross-sectional correlation (unconditionally) which if not rejected will allow us to consider the simpler model under cross-sectional independence as a viable specification. Rejection of the previous hypothesis will indicate that the (unconditional) covariance matrix of the N vector of disturbances is not diagonal making it worth to explore a possible time dependent pattern of the covariance among each pair of units. This can be done in a similar way as outlined previously for the case of the variance. Specifically, we can examine if the cross products of LSDV or LS residuals show a significant autocorrelation pattern. The inclusion of pair specific effects in the covariance process can be decided after testing for individual effects in the AR regression of cross products of residuals.

We need to remark that the previous guidelines can be quite helpful to determine a preliminary specification of the model. However, in order to determine the most appropriate model we need to estimate a few alternative specifications via maximum likelihood and compare the results. At this point, it is important to make sure that all conditional heteroskedasticity has been captured in the estimation. We can accomplish this in two ways. First, we can add additional terms in the conditional variance equation and check for their significance. Second, we can test the squared normalized residuals for any autocorrelation pattern. If significant patterns remain, alternative specifications should be estimated and checked.

[15] This argument is along the lines of Bollerslev (1986) who suggests examining the squared least squares residuals in order to determine the presence of ARCH effects in a time series context.

10.4. Inflation uncertainty in the G7 Countries

Several studies have found using time series GARCH models that inflation uncertainty, measured by the estimated conditional variance, is a significant phenomenon in the G7 and other countries, and that it interacts in various ways with nominal or real variables.[16] In this paper we attempt to characterize the conditional variance–covariance process of inflation in the G7 countries taken as a panel. We also evaluate the hypotheses that (i) higher inflation uncertainty increases average inflation and (ii) higher inflation rates become less predictable. We use monthly observations on inflation rates (π) during the period 1978.2 to 2003.9.[17]

Before proceeding, we evaluate the stationarity of the inflation process. In Table 10.1 we present time series as well as panel unit root tests for inflation. In all cases the regression model for the test includes an intercept. For each individual country, we use the Augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) tests.[18] The results reject the null hypothesis of unit root except in the cases of France and Italy when using the ADF test. At the panel level, both Levin *et al*. (2002) t-star and Im *et al*.'s (2003) t-bar and W (t-bar) tests reject the null of unit root, which enables us to treat this panel as stationary.[19]

10.4.1. Conditional heteroskedasticity and cross-sectional dependence in G7 inflation

In this section we present and briefly discuss the estimation results of various DPDCCV models after performing some preliminary testing following the empirical strategy outlined in Section 10.3. In all cases, we consider an AR (12) specification for the mean equation since we are using seasonally unadjusted monthly data.

First, we test for individual effects in the mean equation. The Wald test statistic (using White/Newey–West's HAC covariance matrix) is $\chi^2_{(6)} =$

[16] See, for example, Caporale and Caporale (2002), Apergis (1999) and Grier and Perry (1996, 1998, 2000) among others. It is also worth mentioning the seminal paper by Robert Engle (1982).

[17] These data are compiled from the International Monetary Fund's (IMF) International Financial Statistics.

[18] For the ADF and PP tests the number of lags was determined by the floor $\{4(T/100)^{5/9}\}$ which gives a value of 5 lags in all cases.

[19] It is important to remark that the alternative hypothesis is not the same. In Levin–Lin–Chu test all cross sections are stationary with the same AR parameter while in the case of Im, Pesaran and Shin the AR parameter is allowed to differ across units and not all individual processes need to be stationary.

Table 10.1. Time series and panel data unit root tests for inflation in the G7 countries

	Augmented Dickey–Fuller	Phillips–Perron $Z(\rho)$	Phillips–Perron $Z(t)$
Time series unit root tests			
Canada	−3.905	−260.545	−13.174
France	−2.131	−71.268	−6.525
Germany	−5.464	−222.100	−12.638
Italy	−2.247	−68.437	−6.282
Japan	−5.521	−219.774	−14.866
U.K.	−3.525	−215.237	−12.291
U.S.	−3.636	−96.258	−7.626
Panel data unit root tests			
Pooled t-star test: (Levin *et al.*, 2002)		−4.79577	(0.0000)
t-bar test: (Im *et al.*, 2003)		−6.227	(0.0000)
$W(t$-bar) test: (Im *et al.*, 2003)		−14.258	(0.0000)

The time series unit root tests correspond to the model with intercept only. For the Augmented Dickey–Fuller (ADF) and the Phillips–Perron (PP) tests, the lag truncation was determined by floor $4(T/100)^{2/9}$. For the ADF and PP $Z(t)$ tests, the approximate 1, 5 and 10 percent critical values are −3.456, −2.878 and −2.570 respectively. For the PP $Z(\rho)$ test the approximate 1 percent critical value is −20.346. For the panel unit root tests, the number of lags for each individual country was also set to floor $4(T/100)^{2/9}$. Numbers in parenthesis are p-values.

2.82, which is not significant at any conventional level and lead us to consider a common intercept in the mean equation.

Secondly, we perform likelihood ratio tests for (unconditional) groupwise heteroskedasticity and cross-sectional correlation obtaining the values of $\chi^2_{(6)} = 255.08$ and $\chi^2_{(21)} = 214.28$ respectively. These tests statistics are highly significant and indicate that the unconditional variance–covariance matrix of disturbances is neither scalar identity nor diagonal.

More explicitly, these results show that there is significant unconditional groupwise heteroskedasticity and cross-sectional correlation. Clearly, the second test suggests that the assumption of cross-sectional independence does not hold in these data.

Next, in order to explore if a significant conditional variance–covariance process exists, we estimate AR (12) regressions using the squared as well as the cross products of the residuals taken from the pooled AR (12) mean inflation regression. For the squared residuals, lag 1 is significant at the

5% while lags 3, 9 and 12 are significant at the 1% level. In the case of the cross products of residuals, lags 1, 3, 6, 7, 9 and 12 are significant at the 1% level.[20]

We also perform simple tests for individual effects in the previous AR (12) regressions. We find that the null of no individual effects in the regression using squared residuals is rejected at the 5% significance level. This result, together with the previous evidence on unconditional groupwise heteroskedasticity, leads us to include individual effects in the conditional variance equation. For the case of cross products of LS residuals, the joint null of no pair specific effects is not rejected pointing to a covariance process with a single intercept.[21]

To summarize, the preliminary testing suggests a dynamic panel model without individual effects in the mean equation for inflation rates (π). For both the conditional variance and conditional covariance processes, a GARCH (1, 1) specification seems to be appropriate given the persistence exhibited by the squares and cross products of the LS residuals respectively. The variance and covariance equations may include individual specific and a single intercept respectively. This DPDCCV model will be estimated and referred to as Model 2. We will also consider a few relevant alternative specifications based on the following benchmark model:

$$\pi_{it} = \mu + \sum_{j=1}^{12} \beta_j \pi_{it-j} + u_{it}, \quad i = 1, \ldots, 7;\; t = 1, \ldots, 296, \tag{10.9}$$

$$\sigma_{it}^2 = \alpha_i + \delta\sigma_{i,t-1}^2 + \gamma u_{i,t-1}^2, \tag{10.10}$$

$$\sigma_{ijt} = \eta_{ij} + \lambda\sigma_{ij,t-1} + \rho u_{i,t-1} u_{j,t-1}. \tag{10.11}$$

This model will be referred to as Model 3. Model 2 is a special case of Model 3 in that $\eta_{ij} = \eta$ in Equation (10.11). We also consider a model with cross-sectional independence, which is defined by Equations (10.9) and (10.10) only. This will be referred to as Model 1. For comparison, two versions of the simple dynamic panel data (DPD) model without GARCH effects are also considered. The first one, which includes country specific effects, is estimated using the LSDV as well as Arellano and Bond's

[20] The results are available upon request.

[21] It is important to note, though, that 9 out of the 21 pair specific coefficients resulted positive and significant at the 10% or less, indicating that a model with pair specific effects in the covariance process may not be discarded.

(1991) GMM estimators.[22] The pooled regression model (common intercept) is estimated by OLS.[23]

In Table 10.2 we report some conventional DPD estimation results. Two issues are worth noting. First, the estimated coefficients for the mean equation are numerically quite close, although the GMM1 estimator is the most efficient (as it would have been expected) and gives a higher number of significant coefficients than the other estimators. Second, when comparing OLS and LSDV results we find that the (implied) values of the log likelihood function are also quite close, which is congruent with the non-rejection results from the Wald test for no individual specific effects reported before.

Given the previous results, we consider that a specification without country specific effects in the mean equation is justified and therefore we use it for the DPDCCV models. The estimation results of these models are shown in Table 10.3. All of them were obtained by MLE. It should be remarked that we estimated 22, 25 and 45 parameters in Models 1, 2 and 3 respectively.

Clearly, the last DPDCCV model (Model 3) outperforms all the other models based on the value of the log-likelihood function. Notice that our specification strategy picked Model 2 rather than Model 3, so that actually estimating several reasonable models is probably important to do in practice. In what follows we use the results of Model 3 to characterize the G7's mean inflation process as well as its associated conditional variance and covariance processes.

According to Model 3, the G7's inflation volatility can be characterized as a significant and quite persistent although stationary GARCH (1, 1) process. Similarly, the results for the covariance equation indicate that this process is also a quite persistent GARCH (1, 1).

We find that all individual specific coefficients in the variance equation are statistically significant at the 1% level. Also, all but two of the pair specific coefficients in the covariance equation are positive and about half of them are statistically significant at the 10% level or less.[24]

Some interesting patterns of individual volatility and cross-sectional dependence among the G7's inflation shocks are worth mentioning. First,

[22] Given that we are dealing with a large T and small N panel we only use the GMM1 estimator after restricting the number of lagged values of the dependent variable to be used as instruments to a maximum of 7. Specifically, we use lags 13th through 19th as instruments. See also Baltagi (2001, pp. 131–136) for details on these estimators.

[23] For both OLS and LSDV we computed standard errors using White/Newey–West's HAC covariance matrix.

[24] These results as well as the ones we referred to in the rest of this section are available upon request.

Table 10.2. Conventional DPD estimation results

DPD Model (individual specific effects): LSDV estimator

Log likelihood $= -5691.96$

Mean:

$$\begin{aligned}\pi_{it} = \mu_i &+ \underset{(6.32)^{***}}{0.176}\,\pi_{it-1} - \underset{(-0.35)}{0.008}\,\pi_{it-2} + \underset{(0.79)}{0.025}\,\pi_{it-3}\\ &+ \underset{(2.05)^{**}}{0.052}\,\pi_{it-4} + \underset{(1.51)}{0.033}\,\pi_{it-5} + \underset{(3.28)^{***}}{0.086}\,\pi_{it-6} + \underset{(2.09)^{**}}{0.046}\,\pi_{it-7}\\ &+ \underset{(0.46)}{0.009}\,\pi_{it-8} + \underset{(0.30)}{0.012}\,\pi_{it-9} - \underset{(-0.58)}{0.013}\,\pi_{it-10} + \underset{(1.83)^{*}}{0.046}\,\pi_{it-11}\\ &+ \underset{(14.39)^{***}}{0.446}\,\pi_{it-12} + \hat{u}_{it}\end{aligned}$$

Variance: $\sigma_{it}^2 = 14.38$

Covariance: $\sigma_{ijt} = 0$

DPD Model (individual specific effects): Arellano–Bond GMM1 estimator

Mean:

$$\begin{aligned}\pi_{it} = \mu_i &+ \underset{(33.32)^{***}}{0.173}\,\pi_{it-1} - \underset{(-0.77)}{0.004}\,\pi_{it-2} + \underset{(3.54)^{***}}{0.019}\,\pi_{it-3}\\ &+ \underset{(10.25)^{***}}{0.054}\,\pi_{it-4} + \underset{(6.32)^{***}}{0.033}\,\pi_{it-5} + \underset{(16.29)^{***}}{0.086}\,\pi_{it-6}\\ &+ \underset{(8.77)^{***}}{0.046}\,\pi_{it-7} + \underset{(0.69)}{0.004}\,\pi_{it-8} + \underset{(2.79)^{**}}{0.015}\,\pi_{it-9} - \underset{(1.54)}{0.008}\,\pi_{it-10}\\ &+ \underset{(9.15)^{***}}{0.048}\,\pi_{it-11} + \underset{(86.96)^{***}}{0.446}\,\pi_{it-12} + \hat{u}_{it}\end{aligned}$$

Variance: $\sigma_{it}^2 = 13.72$

Covariance: $\sigma_{ijt} = 0$

DPD Model (common intercept): OLS estimator

Log likelihood $= -5692.45$

Mean:

$$\begin{aligned}\pi_{it} = \underset{(1.30)}{0.172} &+ \underset{(6.35)^{***}}{0.177}\,\pi_{it-1} - \underset{(-0.31)}{0.007}\,\pi_{it-2} + \underset{(0.83)}{0.026}\,\pi_{it-3}\\ &+ \underset{(2.07)^{**}}{0.052}\,\pi_{it-4} + \underset{(1.55)}{0.034}\,\pi_{it-5} + \underset{(3.34)^{***}}{0.087}\,\pi_{it-6} + \underset{(2.14)^{**}}{0.047}\,\pi_{it-7}\\ &+ \underset{(0.50)}{0.010}\,\pi_{it-8} + \underset{(0.32)}{0.012}\,\pi_{it-9} - \underset{(-0.54)}{0.012}\,\pi_{it-10} + \underset{(1.85)^{*}}{0.047}\,\pi_{it-11}\\ &+ \underset{(14.50)^{***}}{0.447}\,\pi_{it-12} + \hat{u}_{it}\end{aligned}$$

Variance: $\sigma_{it}^2 = 14.25$

Covariance: $\sigma_{ijt} = 0$

For each model we show the estimated mean equation followed by the estimated (or implied) equations for the conditional variance and covariance processes. Values in parenthesis are t-ratios. The t-ratios for the OLS and LSDV estimators are based on White/Newey–West's HAC standard errors. For the GMM1 estimator the number of lagged values of the dependent variable to be used as instruments is restricted to a maximum of 7. Specifically, we use lags 13th through 19th as instruments.

*indicate significance level of 10%.

**indicate significance level of 5%.

***indicate significance level of 1%.

Table 10.3. ***Estimation results for the DPDCCV model***

DPDCCV Model 1 (conditional variance only): MLE estimator

Log likelihood $= -5466.22$

Mean:

$$\pi_{it} = \underset{(3.13)^{***}}{0.331} + \underset{(8.69)^{***}}{0.193}\,\pi_{it-1} - \underset{(-1.62)}{0.036}\,\pi_{it-2} + \underset{(1.16)}{0.025}\,\pi_{it-3}$$
$$+ \underset{(1.64)}{0.036}\,\pi_{it-4} + \underset{(0.19)}{0.004}\,\pi_{it-5} + \underset{(3.38)^{***}}{0.071}\,\pi_{it-6} + \underset{(2.84)^{***}}{0.060}\,\pi_{it-7}$$
$$+ \underset{(1.35)}{0.027}\,\pi_{it-8} - \underset{(-1.33)}{0.027}\,\pi_{it-9} + \underset{(1.14)}{0.024}\,\pi_{it-10} + \underset{(2.72)^{***}}{0.056}\,\pi_{it-11}$$
$$+ \underset{(21.94)^{***}}{0.434}\,\pi_{it-12} + \hat{u}_{it}$$

Variance:

$$\sigma_{it}^2 = \alpha_i + \underset{(29.61)^{***}}{0.769}\,\sigma_{i,t-1}^2 + \underset{(6.90)^{***}}{0.148}\,u_{i,t-1}^2$$

Covariance: $\sigma_{ijt} = 0$

DPDCCV Model 2 (conditional variance and covariance): MLE estimator

Log likelihood $= -5355.61$

Mean:

$$\pi_{it} = \underset{(2.93)^{***}}{0.367} + \underset{(6.71)^{***}}{0.153}\,\pi_{it-1} - \underset{(-1.71)^{*}}{0.039}\,\pi_{it-2} + \underset{(0.53)}{0.012}\,\pi_{it-3}$$
$$+ \underset{(0.95)}{0.021}\,\pi_{it-4} + \underset{(0.81)}{0.018}\,\pi_{it-5} + \underset{(3.64)^{***}}{0.078}\,\pi_{it-6} + \underset{(2.91)^{***}}{0.062}\,\pi_{it-7}$$
$$+ \underset{(1.06)}{0.022}\,\pi_{it-8} - \underset{(-1.21)}{0.026}\,\pi_{it-9} + \underset{(1.60)}{0.034}\,\pi_{it-10} + \underset{(2.49)^{**}}{0.052}\,\pi_{it-11}$$
$$+ \underset{(21.27)^{***}}{0.443}\,\pi_{it-12} + \hat{u}_{it}$$

Variance:

$$\sigma_{it}^2 = \alpha_i + \underset{(48.95)^{***}}{0.884}\,\sigma_{i,t-1}^2 + \underset{(5.68)^{***}}{0.072}\,u_{i,t-1}^2$$

Covariance:

$$\sigma_{ijt} = \underset{(2.66)^{***}}{0.072} + \underset{(26.08)^{***}}{0.877}\,\sigma_{ijt-1} + \underset{(3.86)^{***}}{0.037}\,u_{i,t-1}u_{j,t-1}$$

(*Continued on next page*)

our results suggest that Italy, France and USA have the lowest levels of unconditional volatility. Second, the USA has relatively high and significant positive cross-sectional dependence with Canada and to a lesser extent with France, Germany and Italy. Third, Japan's inflation shocks do not seem to be correlated with any of the other G7 countries. Fourth, the three biggest European economies, namely France, Germany and UK show a relatively significant pattern of positive cross-sectional dependence.

We also find some interesting patterns for the conditional volatility processes. For example while in most G7's the volatility levels appear to be lower at the end of the sample compared with those experienced in the eighties, this does not appear to be the case for Canada and Germany. Also, the volatility levels appear to have been rising in the last two years of the sample in the cases of Canada, France, Germany, and the USA.

Table 10.3. (Continued)

DPDCCV Model 3 (conditional variance and covariance): MLE estimator

Log likelihood = -5328.08

Mean:

$$\pi_{it} = \underset{(3.23)^{***}}{0.407} + \underset{(6.69)^{***}}{0.154}\,\pi_{it-1} - \underset{(-1.46)}{0.033}\,\pi_{it-2} + \underset{(0.87)}{0.020}\,\pi_{it-3}$$

$$+\underset{(0.92)}{0.020}\,\pi_{it-4} + \underset{(1.00)}{0.022}\,\pi_{it-5} + \underset{(3.50)^{***}}{0.075}\,\pi_{it-6} + \underset{(2.88)^{***}}{0.061}\,\pi_{it-7}$$

$$+\underset{(0.77)}{0.016}\,\pi_{it-8} - \underset{(-1.33)}{0.029}\,\pi_{it-9} + \underset{(1.56)}{0.033}\,\pi_{it-10} + \underset{(2.50)^{**}}{0.052}\,\pi_{it-11}$$

$$+\underset{(20.71)^{***}}{0.433}\,\pi_{it-12} + \hat{u}_{it}$$

Variance:

$$\sigma^2_{it} = \alpha_i + \underset{(44.06)^{***}}{0.882}\,\sigma^2_{i,t-1} + \underset{(5.26)^{***}}{0.069}\,u^2_{i,t-1}$$

Covariance:

$$\sigma_{ijt} = \eta_{ij} + \underset{(11.96)^{***}}{0.806}\,\sigma_{ijt-1} + \underset{(2.98)^{***}}{0.034}\,u_{i,t-1}u_{j,t-1}$$

For each model we show the estimated mean equation followed by the estimated (or implied) equations for the conditional variance and covariance processes. Values in parenthesis are t-ratios. All DPDCCV models were estimated by direct maximization of the log-likelihood function using numerical methods.

*indicate significance level of 10%.

**indicate significance level of 5%.

***indicate significance level of 1%.

We have also calculated the implied conditional cross correlations between the USA and the other G7 countries and between France, Germany and Italy. The dependence of USA with Canada, France and Italy seems to have increased over time. On the other hand, the process does not seem to exhibit a clear pattern over time in the case of the three biggest European economies.

10.4.2. The interrelationship between average inflation and inflation uncertainty

One advantage of our DPDCCV model over conventional DPD models and their associated estimation methods, including GMM, is that it allows us to directly test some interesting hypotheses about the interrelationship between average inflation and inflation uncertainty. The most famous of these, that higher average inflation is less predictable, is due to Friedman (1977) and was formalized by Ball (1992). We can test this hypothesis for the G7 countries by including lagged inflation as a regressor in our conditional variance equation.

It has also been argued that increased inflation uncertainty can affect the average inflation rate. The theoretical justification for this hypothesis is given in Cukierman and Meltzer (1986) and Cukierman (1992) where it

Table 10.4. Estimation results for the DPDCCV model with variance effects in the conditional mean and lagged inflation in the conditional variance

DPDCCV Model 4 (conditional variance and covariance): MLE estimator

Log likelihood = −5306.03

Mean:

$$\pi_{it} = \underset{(3.22)^{*}}{0.643} + \underset{(6.89)^{***}}{0.156}\,\pi_{it-1} - \underset{(-1.12)}{0.025}\,\pi_{it-2} + \underset{(0.97)}{0.021}\,\pi_{it-3}$$

$$+ \underset{(1.17)}{0.026}\,\pi_{it-4} + \underset{(1.42)}{0.031}\,\pi_{it-5} + \underset{(4.02)^{***}}{0.086}\,\pi_{it-6} + \underset{(2.92)^{***}}{0.061}\,\pi_{it-7}$$

$$+ \underset{(0.84)}{0.018}\,\pi_{it-8} - \underset{(-1.03)}{0.022}\,\pi_{it-9} + \underset{(1.23)}{0.026}\,\pi_{it-10} + \underset{(2.78)^{***}}{0.057}\,\pi_{it-11}$$

$$+ \underset{(20.92)^{***}}{0.438}\,\pi_{it-12} - \underset{(-1.72)^{**}}{0.117}\,\sigma_{it} + \hat{u}_{it}$$

Variance:

$$\sigma_{it}^2 = \alpha_i + \underset{(39.24)^{***}}{0.867}\,\sigma_{i,t-1}^2 + \underset{(4.53)^{***}}{0.050}\,u_{i,t-1}^2 + \underset{(4.71)^{***}}{0.092}\,\pi_{i,t-1}$$

Covariance:

$$\sigma_{ijt} = \eta_{ij} + \underset{(17.84)^{***}}{0.855}\,\sigma_{ijt-1} + \underset{(3.02)^{***}}{0.031}\,u_{i,t-1}u_{j,t-1}$$

This model has been estimated by direct maximization of the log-likelihood function by numerical methods. We show the estimated equations for the conditional mean, variance and covariance processes. Values in parenthesis are t-ratios.

*indicate significance level of 10%.

**indicate significance level of 5%.

***indicate significance level of 1%.

is shown that increases in inflation uncertainty increase the policy maker's incentive to create inflation surprises, thus producing a higher average inflation rate. In order to evaluate the previous hypothesis we simply include the conditional variance as an additional regressor in the mean equation. To conduct these tests, we alter Equations (10.9) and (10.10) as shown below and call the resulting system Model 4 (we continue to use Equation (10.11) for the covariance process).

$$\pi_{it} = \mu + \sum_{j=1}^{12} \beta_j \pi_{it-j} + \kappa\sigma_{it} + u_{it}, \quad i = 1, \ldots, 7;\ t = 1, \ldots, 296, \tag{10.9a}$$

$$\sigma_{it}^2 = \alpha_i + \delta\sigma_{i,t-1}^2 + \gamma u_{i,t-1}^2 + \psi\pi_{i,t-1}. \tag{10.10a}$$

A positive and significant value for the parameter κ supports the Cukierman and Meltzer hypothesis that inflation volatility raises average inflation, while a positive and significant value for the parameter ψ supports the Friedman–Ball hypothesis that higher inflation is more volatile.

The results are shown in the Table 10.4. As it can be seen, the parameter ψ is positive and highly statistically significant, indicating that higher inflation rates do become less predictable as argued by Friedman. On the other hand, we find that the parameter κ is significant at the 5% level although its sign is negative, which clearly rejects the hypothesis that higher inflation uncertainty produces higher average inflation rates. This negative sign actually supports previous findings by Holland (1995) for the USA and by Grier and Perry (1998) for the USA and Germany. These authors argue that if inflation uncertainty has deleterious real effects that central banks dislike and if higher average inflation raises uncertainty (as we have found here) then the Central Bank has a stabilization motive to reduce uncertainty by reducing average inflation. In our G7 panel we find the stabilization motive dominates any potentially opportunistic Central Bank behavior.

Overall, when comparing Model 3 in Table 10.3 with Model 4 in Table 10.4 by means of a likelihood ratio test we find that the later outperforms to the former and lead us to conclude that (i) higher inflation rates are less predictable and (ii) higher inflation uncertainty has been associated with lower average inflation rates.

10.5. Conclusion

In this paper we have specified a model, (DPDCCV), which accounts for conditional heteroskedasticity and cross-sectional correlation within a panel data framework, an issue that has not yet been addressed in the panel data literature. We have also outlined a methodology to identify these phenomena, which could be useful for empirical research.

The DPDCCV model has been applied to a panel of monthly inflation rates for the G7, over the period 1978.2–2003.9, showing that there exist highly persistent patterns of volatility as well as cross-sectional dependence. Further, we have found that higher inflation rates become less predictable. Also, we have found that the hypothesis that higher inflation uncertainty produces higher average inflation rates is not supported in these data. On the contrary, we find that this relationship is negative indicating that Central Banks dislike inflation uncertainty.

Although the model formulated here is practical for small N and large T panels, it is especially relevant due to the following 4 factors: (1) The rapid growth of empirical panel research on macroeconomic and financial issues, (2) The ubiquity of conditional heteroskedasticity in macroeconomic and financial data, (3) The potential extreme inefficiency of estimators that do not account for these phenomena, and (4) The rapid growth of multivariate GARCH models outside the panel data literature. Further work, particularly theoretical, to account for these phenomena in a more general panel setting is certainly necessary.

Acknowledgements

The authors would like to thank Paul Evans, Tim Bollerslev, Badi Baltagi and two anonymous referees for valuable comments. Earlier versions of this paper were presented at the 2001 North American Summer Meetings and 2003 Latin American Meetings of the Econometric Society, the 8th Annual Meeting of LACEA, and the 11th International Conference on Panel Data. We thank all comments from participants at these conferences. Any remaining mistakes are only the authors' responsibility.

References

Apergis, N. (1999), "Inflation uncertainty and money demand: evidence from a monetary regime change and the case of Greece", *International Economic Journal*, Vol. 13 (2), pp. 21–23.

Arellano, M. (1987), "Computing robust standard errors for within-groups estimators", *Oxford Bulletin of Economics and Statistics*, Vol. 49, pp. 431–434.

Arellano, M., Bond, S. (1991), "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations", *Review of Economic Studies*, Vol. 58, pp. 277–297.

Ball, L. (1992), "Why does high inflation raise inflation uncertainty?", *Journal of Monetary Economics*, Vol. 29, pp. 371–388.

Baltagi, B.H. (2001), *Econometric Analysis of Panel Data*, 2nd ed., John Wiley.

Bauwens, L., Laurent, S., Rombouts, J.V.K. (2003), "Multivariate GARCH models: a survey", CORE Discussion Paper 2003/31.

Bollerslev, T. (1986), "Generalized autoregressive conditional heteroskedasticity", *Journal of Econometrics*, Vol. 31, pp. 307–327.

Bollerslev, T. (1990), "Modeling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model", *Review of Economics and Statistics*, Vol. 72, pp. 498–505.

Bollerslev, T., Wooldridge, J. (1992), "Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances", *Econometric Reviews*, Vol. 11 (2), pp. 143–172.

Bollerslev, T., Engle, R., Wooldridge, J. (1988), "A capital asset pricing model with time varying conditional covariances", *Journal of Political Economy*, Vol. 96, pp. 116–131.

Bollerslev, T., Chou, R.Y., Kroner, K. (1992), "ARCH modeling in finance: a review of the theory and empirical evidence", *Journal of Econometrics*, Vol. 52, pp. 5–59.

Brockwell, P.J., Davis, R.A. (1991), *Time Series: Theory and Methods*, 2nd ed., Springer-Verlag.

Caporale, B., Caporale, T. (2002), "Asymmetric effects of inflation shocks on inflation uncertainty", *Atlantic Economic Journal*, Vol. 30 (4), pp. 385–388.

Cukierman, A. (1992), *Central Bank Strategy, Credibility and Independence*, MIT Press.

Cukierman, A., Meltzer, A. (1986), "A theory of ambiguity, credibility and inflation under discretion and asymmetric information", *Econometrica*, Vol. 50, pp. 1099–1128.

Davidson, R., McKinnon, J.G. (1993), *Estimation and Inference in Econometrics*, Oxford University Press.

Engle, R. (1982), “Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation”, *Econometrica,* Vol. 50, pp. 987–1007.

Engle, R. (2002), “Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models”, *Journal of Business and Economic Statistics,* Vol. 20, pp. 339–350.

Engle, R., Lillien, D., Robbins, R. (1987), “Estimating time varying risk premia in the term structure: the ARCH-M model”, *Econometrica,* Vol. 55, pp. 391–407.

Friedman, M. (1977), “Nobel lecture: inflation & unemployment”, *Journal of Political Economy,* Vol. 85, pp. 451–472.

Grier, K., Perry, M.J. (1996), “Inflation uncertainty and relative price dispersion”, *Journal of Monetary Economics,* Vol. 38 (2), pp. 391–405.

Grier, K., Perry, M.J. (1998), “On inflation and inflation uncertainty in the G7 countries”, *Journal of International Money and Finance,* Vol. 17 (4), pp. 671–689.

Grier, K., Perry, M.J. (2000), “The effects of real and nominal uncertainty on inflation and output growth: some GARCH-M evidence”, *Journal of Applied Econometrics,* Vol. 15, pp. 45–58.

Harvey, A. (1990), *The Econometric Analysis of Time Series*, 2nd ed., MIT Press.

Holland, A.S. (1995), “Inflation and uncertainty: tests for temporal ordering”, *Journal of Money, Credit and Banking,* Vol. 27, pp. 827–837.

Hsiao, C. (2003), *Analysis of Panel Data*, 2nd ed., Cambridge University Press.

Im, K.S., Pesaran, M.H., Shin, Y. (2003), “Testing for unit roots in heterogeneous panels”, *Journal of Econometrics,* Vol. 115, pp. 53–74.

Kiviet, J.F. (1995), “On bias, inconsistency, and efficiency of various estimators in dynamic panel data models”, *Journal of Econometrics,* Vol. 68, pp. 53–78.

Ledoit, O., Santa-Clara, P., Wolf, M. (2003), “Flexible multivariate GARCH models with an application to international stock markets”, *Review of Economics and Statistics,* Vol. 85 (3), pp. 735–747.

Levin, A., Lin, C.-F., Chu, C.-S.J. (2002), “Unit root tests in panel data: asymptotic and finite-samples properties”, *Journal of Econometrics,* Vol. 108, pp. 1–24.

Newey, W., West, K. (1987), “A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix”, *Econometrica,* Vol. 55, pp. 703–708.

Phillips, P.C.B., Sul, D. (2003), “Dynamic panel estimation and homogeneity testing under cross section dependence”, *Econometrics Journal,* Vol. 6, pp. 217–259.

White, H. (1980), “A heteroskedasticity-consistent covariance estimator and a direct test for heteroskedasticity”, *Econometrica,* Vol. 48, pp. 817–838.

This page intentionally left blank

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74011-1

CHAPTER 11

The Dynamics of Exports and Productivity at the Plant Level: A Panel Data Error Correction Model (ECM) Approach

Mahmut Yasar[*,a], Carl H. Nelson[b] and Roderick M. Rejesus[c]

[a]Department of Economics, Emory University, 306C Rich Building, Atlanta, GA 30322, USA
E-mail address: myasar@emory.edu
[b]Department of Agricultural & Consumer Economics, University of Illinois at Urbana-Champaign, 1301 W. Gregory Drive, Urbana, IL 61801, USA
E-mail address: chnelson@uiuc.edu
[c]Department of Agricultural & Applied Economics, Texas Tech University, Box 42123, Lubbock, TX 79409-2132, USA
E-mail address: roderick.rejesus@ttu.edu

Abstract

This article examines the short-run and long-run dynamics of the export-productivity relationship for Turkish manufacturing industries. We use an error correction model (ECM) estimated using a system Generalized Method of Moments (GMM) estimator to achieve this objective. Our results suggest that permanent productivity shocks generate larger long-run export level responses, as compared to long-run productivity responses from permanent export shocks. This result suggests that industrial policy should be geared toward permanent improvements in plant-productivity in order to have sustainable long-run export and economic growth.

Keywords: Europe, exporting, long-run dynamics, productivity, short-run dynamics, Turkey

JEL classifications: F10, F14, D21, L60

11.1. Introduction

There have been a multitude of recent studies that examine the relationship between exports and productivity using panel data. For example, studies

[*] Corresponding author.

by Bernard and Jensen (1998, 1999) on the United States; Aw *et al.* (1998) on Taiwan and Korea; Clerides *et al.* (1998) on Colombia, Mexico, and Morocco; Kraay (1997) on China; Wagner (2002) on Germany; and Girma *et al.* (2003) on the U.K., have all examined the exports-productivity link using panel data sets. However, all of these studies have not explicitly examined the issues related to the short-run and long-run dynamics of the relationship between exports and productivity. This is an important issue to investigate since both the international trade and the endogenous growth literatures have developed theories that show the existence and importance of discerning the long-run relationship between exports and productivity (Dodaro, 1991).

In using panel data sets to empirically determine the dynamics of the export and productivity relationship, there are two main issues that one has to deal with. First, it is likely that exports and productivity are correlated with the current realization of the unobserved firm- or plant-specific effects (Marschak and Andrews, 1944). For example, unobserved factors such as the managerial ability of the CEO or effectiveness of research and development efforts could have an impact on the firm's productivity and/or exporting status. Second, exporting and productivity tend to be highly persistent over time and are typically jointly determined (i.e. they are endogenous). In such a case, adjustments to unobserved shocks may not be immediate but may occur with some delay and the strict exogeneity of the explanatory variable(s) conditional on unobserved plant characteristics will not be satisfied. From standard econometric theory, if these two issues are not addressed, then the long-run and short-run dynamics of the export and productivity relationship will not be consistently estimated and standard inference procedures may be invalid. Kraay (1997) addressed the two issues above by employing a first-differenced GMM estimator. The difficulty of unobserved firm-specific effects was handled by working with a first-differenced specification, while the presence of lagged performance and an appropriate choice of instruments address the issue of persistence and endogeneity. However, the short-run and long-run dynamics of the export-productivity relationship has not been explicitly examined by the previous studies. The exception is Chao and Buongiorno (2002) where they empirically explored the long run multipliers associated with the export-productivity relationship.

The main objective of our paper, therefore, is to determine the short-run adjustments and the long-run relationship of exports and productivity for the case of two Turkish manufacturing industries – the textile and apparel (T&A) industry and the motor vehicle and parts (MV&P) industry. An error correction model (ECM) estimated using a system Generalized Method of Moments (GMM) estimator is used to achieve this objective, while at the same time address the two main difficulties in dynamic panel analysis explained above – unobserved firm specific effects and persistence/endogeneity. Investigating the short-run and long-run dynamics of the export-productivity link is important since this kind of analysis will yield important insights that could guide industrial policy in a low-middle income economy like Turkey. Turkey is a pertinent case for studying the export-productivity link because after trade liberalization in the early 1980s, exports grew at a high rate but have steadily declined ever since. Is this because external policies to promote export growth do not result in sustainable long-run adjustments in productivity to maintain this growth? Uncovering the answer to this type of question would help Turkey set policies that would spur economic growth in the future.

The remainder of the paper is organized as follows. The conceptual foundations for the long-run relationship between exports and productivity, as well as the theoretical explanations for bi-directional causation, are presented in Section 11.2. Section 11.3 discusses the empirical approach and the data used for the analysis. Section 11.4 reports the empirical results and Section 11.5 concludes.

11.2. Conceptual framework

In this section, we explain the conceptual issues that link exports and productivity in the long-run, as well as the plausibility of a bi-directional causality between the two variables. There are two strands of literature that support opposing directional theories of causation. First, the endogenous growth theory posits that the direction of causation flows from exports to productivity and that the long-run relationship is based on this causation. In contrast, the international trade literature suggests that the long-run relationship between exports and productivity is where the direction of causation flows from productivity to exports. We discuss these two contrasting theoretical explanation for the long-run link between exports and productivity in turn.

The causation from exports to productivity is more popularly known in the literature as the learning-by-exporting explanation for the long-run link between exports and productivity. Various studies in the endogenous growth literature argue that exports enhance productivity through innovation (Grossman and Helpman, 1991; Rivera-Batiz and Romer, 1991), technology transfer and adoption from leading nations (Barro and Sala-I-Martin, 1995; Parente and Prescott, 1994), and learning-by-doing gains (Lucas, 1988; Clerides *et al.*, 1998). The innovation argument is where firms are forced to continually improve technology and product standards to compete in the international market. The technological and learning-by-doing gains arise because of the exposure of exporting firms to cutting-edge technology and managerial skills from their international counterparts. Economies of scale from operating in several international markets are also often cited as one other explanation for the learning-by-exporting hypothesis.

On the other hand, the causation from productivity to exports is another theoretical explanation put forward to explain the long-run link between exports and productivity. This line of reasoning is known in the literature as the self-selection hypothesis. According to this hypothesis, firms that have higher productivity are more likely to cope with sunk costs associated with entry to the export market and are more likely to survive the more competitive international markets. This is in line with the findings from international trade theory that firms self-select themselves into export markets (Roberts and Tybout, 1997; Clerides *et al.*, 1998; Bernard and Jensen, 2001). Furthermore, this type of explanation is more in line with the traditional Hecksher–Ohlin notions that increased factor endowments and improved production technologies influence the patterns of trade of specific products (Pugel, 2004). Plants that increase their level of factor endowments or improve production technologies enhance their comparative advantage (relative to plants in other countries) and thus will eventually be able to enter/survive the international market (Dodaro, 1991).

The explanations above provide the conceptual foundations for the link between exports and productivity. Note that these two explanations are not necessarily mutually exclusive and both these theoretical explanations can shape the export-productivity relationship at the same time. Therefore, the empirical question of interest is really to know which explanation is more prominent or dominant for different industries and countries using micro-level data. Furthermore, only an empirical analysis can provide insights as to whether the conceptual explanations above are more prevalent in the short-run or in the long-run. This paper contributes to the literature in this regard.

11.3. Empirical approach and the data

We examine the short-run and long-run dynamics of the exporting and productivity relationship using a generalized one-step ECM estimated using a system GMM estimator. This approach to analyzing short-run and long-run dynamics using panel data is similar to the approach taken by Bond *et al.* (1997, 1999); and Mairesse *et al.* (1999).[1]

11.3.1. The error correction model

We begin with the following autoregressive-distributed lag model:

$$\begin{aligned}\ln y_{i,t} &= \alpha_1 \ln y_{i,t-1} + \alpha_2 \ln y_{i,t-2} + \beta_0 \ln x_{i,t} + \beta_1 \ln x_{i,t-1} \\ &\quad + \beta_2 \ln x_{i,t-2} + \psi_t + v_{it},\end{aligned} \tag{11.1}$$

where $v_{i,t} = \varepsilon_i + u_{i,t}$ and $i = 1, \ldots, N, t = 1, \ldots, T$. Furthermore, i represents the cross-sectional units; t represents the time periods; $y_{i,t}$ is our productivity measure (e.g., total factor productivity or labor productivity); $x_{i,t}$ is the amount of exports;[2] ψ_t is the time-specific effect; and assuming fixed effects, the cross section error term, $v_{i,t}$, contains the following two effects: (1) the unobserved time-invariant, plant-specific effect, ε_i, and (2) a stochastic error term, $u_{i,t}$, varying across time and cross section. The time-specific effect is included to capture aggregate shocks, which may appear in any year. The plant-specific effect, ε_i, is included to capture plant-specific differences such as managerial ability, geographical location, and other unobserved factors. The unobserved plant-specific effect, ε_i, is correlated with the explanatory variables, but not with the changes in the explanatory variables.

The autoregressive-distributed lag model specification is appropriate if the short-run relationship between exporting and productivity is the only object of interest. However, it does not allow for a distinction between the long and short-run effects. We incorporate this distinction into our model by using an error correction specification of the dynamic panel model. This

[1] For more details about the generalized one-step ECM in a time-series context, see Davidson *et al.* (1978) and Banerjee *et al.* (1990, 1993, 1998). In the panel data context, one may also examine related studies by Levin and Lin (1993), Im *et al.* (1997), and Binder *et al.* (2005) where the issues of cointegration and unit roots in panel vector autoregressions are discussed. For more information about the system GMM estimator, in general, see Arellano and Bover (1995) and Blundell and Bond (1998, 2000). A more detailed discussion of the system GMM procedure in the context of this study (i.e. a panel-ECM) is presented in Section 11.3.2 of this article.

[2] In the results section, we refer to this variable as EXP. This is just the amount of exports valued in Turkish Liras. We also deflated this value by the appropriate index so as to be comparable across time. Furthermore, without loss of generality, we only include the amount of exports as the sole explanatory variable in our exposition here. But the model is generalizable to multiple continuous explanatory variables. In addition, as we discuss in the results section below, the ECM model could also be specified with exports as the dependent variable and the productivity as the independent variable. The choice of specification will depend on the assumption about the direction of causation.

error correction specification is a linear transformation of the variables in Equation (11.1), which provides an explicit link between the short-run effects and long run effects (Banerjee *et al.*, 1993, 1998):

$$\begin{aligned}\Delta \ln y_{i,t} &= (\alpha_1 - 1)\Delta \ln y_{i,t-1} + \beta_0 \Delta \ln x_{i,t} + (\beta_0 + \beta_1)\Delta \ln x_{i,t-1} \\ &\quad + \eta(\ln y_{i,t-2} - \ln x_{i,t-2}) + \theta \ln x_{i,t-2} + \psi_t + v_{it}, \\ &\text{where: } \theta = \beta_0 + \beta_1 + \beta_2 + \alpha_2 + \alpha_1 - 1 \text{ and } \eta = \alpha_2 + \alpha_1 - 1.\end{aligned} \tag{11.2}$$

For non-zero values of η this is an error correction model (ECM). The coefficient on the error correction term, $(\ln y_{i,t-2} - \ln x_{i,t-2})$ gives the adjustment rate at which the gap between exporting and productivity is closed. If η is negative and significant, then we conclude that the relationship between exporting and productivity exists in the long-run and the error correction mechanism induces the productivity adjustments to close the gap with respect to the long run relationship between productivity and exporting. Productivity could deviate from the long-run equilibrium relationship due to certain shocks in the short-run, but it eventually converges to the equilibrium in the absence of the shocks in subsequent periods. In such a framework, the long-run productivity dynamics are driven by both the changes in the amount of exports and by the stable nature of the long-run equilibrium.

In this specification, if the coefficient on the error correction term is significantly less than zero, one can conclude that the change in productivity in period t is equal to the change in the exports in period t and the correction for the change between productivity and its equilibrium value in period $t - 1$. If productivity is greater than its equilibrium level, it must decrease for the model to approach equilibrium, and vice-versa. If the model is in equilibrium in period $t - 1$, the error correction term does not influence the change in exports in period t. In this case, the change in the productivity in period t is equal to the change in the independent variable in period t. The error-correcting model allows us to describe the adjustment of the deviation from the long-run relationship between exporting and productivity. In this specification, the first three terms (lagged growth rate of productivity, the contemporaneous and the one-period lagged growth of exports) capture the short-run dynamics and the last two terms (error correction and the lagged level of independent variable) provide a framework to test the long-run relationship between productivity and exports.

In general, a long-run multiplier (ϕ) is typically estimated separately and used to form the error correction term $(\ln y_{i,t-2} - \phi \ln x_{i,t-2})$. With the use of $(\ln y_{i,t-2} - \ln x_{i,t-2})$, the long-run relationship is restricted to be homogeneous (Banerjee *et al.*, 1990, 1993). That is, the implied coefficient of $\phi = 1$ indicates a proportional long-run relationship between y and x. We also use the error correction term of the form $(\ln y_{i,t-2} - \phi \ln x_{i,t-2})$ to avoid this restrictive homogeneity assumption. Thus, in our formulation of the error correction model, we can interpret the coefficient η directly as adjustments to disequilibrium although the true equilibrium is given by $(\ln y_{i,t-2} - \phi \ln x_{i,t-2})$ instead of $(\ln y_{i,t-2} - \ln x_{i,t-2})$. Using this form of the error correction term also allows us to calculate the true long-run relationship between exporting and productivity, which can be written as $1 - (\hat{\theta}/\hat{\eta})$. The error correction specification of the autoregressive distributed lag model that we used here then permits us to directly calculate and analyze the short-run and long-run dynamics of the productivity and exporting relationship.

11.3.2. The system GMM estimation procedure

For consistent and efficient parameter estimates of the panel data error correction model specified in Equation (11.2), we apply the system GMM approach proposed by Arellano and Bover (1995) and Blundell and Bond (1998, 2000). This estimation procedure is especially appropriate when: (i) N is large, but T is small; (ii) the explanatory variables are endogenous; and (iii) unobserved plant-specific effects are correlated with other regressors. Under the assumptions that u_{it} are serially uncorrelated and that the explanatory variables are endogenous, Arellano and Bond (1991) showed that the following moment conditions hold for the equations in first differences:[3]

$$\begin{aligned} &E(\Delta u_{i,t} y_{i,t-r}) = 0; \qquad E(\Delta u_{i,t} x_{i,t-r}) = 0; \\ &\quad \text{where } r = 2, \ldots, t-1 \text{ and } t = 3, \ldots, T. \end{aligned} \tag{11.3}$$

Therefore, the lagged values of endogenous variables dated $t - 2$ and earlier are valid instruments for the equations in first differences.

As a result, Arellano and Bond (1991) showed that the first-differenced GMM estimator method results in a significant efficiency gain compared to the Anderson and Hsiao (1981) estimator.[4] However, in the context of the model specification in (11.2), there are two possible problems with the use of the first differenced GMM estimator. First, the plant-specific

[3] We assume that the explanatory variable is endogenous, i.e. $E(x_{ir}u_{it}) = 0$ for $r = 1, \ldots, t-1$; $t = 2, \ldots, T$; and $E(x_{ir}u_{it}) \neq 0$ for $r = s, \ldots, T$; $s = 2, \ldots, T$. The resulting moment conditions, and thus instruments, would be different if one assumes that the explanatory variables are strictly exogenous or weakly exogenous (see Blundell *et al.*, 2000).

[4] See Baltagi (2001) for a more detailed discussion of this issue.

effect is eliminated, so that one cannot examine the cross-plant relationship between the variables of interest, in our case exports and productivity. Second, when the lagged values of the series are weakly correlated with the first-differences, it can yield parameter estimates that suffer from large finite sample bias because of weak instruments. When the individual series for the dependent and independent variable are highly persistent, and when T is small, the problem is more severe.

Arellano and Bover (1995), however, noted that if the initial condition, X_{i1}, satisfies the stationarity restriction $E(\Delta X_{i2}\varepsilon_i) = 0$, then ΔX_{it} will be correlated with ε_i if and only if ΔX_{i2} is correlated with ε_i. The resulting assumption is that although there is a correlation between the level of right-hand side variables, X_{it}, and the plant-specific effect, ε_i, no such correlation exists between the differences of right-hand side variables, ΔX_{it}, and the plant-specific effect, ε_i. This additional assumption gives rise to the level equation estimator, which exploits more moment conditions. Lagged differences of explanatory variables, ΔX_{it-r}, are used as additional instruments for the equations in levels, when X_{it} is mean stationary.

Blundell and Bond (1998) showed that the lagged differences of the dependent variable, in addition to the lagged differences of the explanatory variables, are proper instruments for the regression in the level equation as long as the initial conditions, y_{i1}, satisfy the stationary restriction, $E(\Delta Y_{i2}\varepsilon_i) = 0$. Thus, when both ΔX_{it} and ΔY_{it} are uncorrelated with ε_i, both lagged differences of explanatory variables, ΔX_{it-r} and lagged differences of dependent variable, ΔY_{it-r}, are valid instruments for the equations in levels. Furthermore, Blundell and Bond (1998) show that the moment conditions defined for the first-differenced equation can be combined with the moment conditions defined for the level equation to estimate a system GMM. When the explanatory variable is treated as endogenous, the GMM system estimator utilizes the following moment conditions:

$$E(\Delta u_{i,t} y_{i,t-r}) = 0; \qquad E(\Delta u_{i,t} x_{i,t-r}) = 0;$$
$$\text{where } r = 2, \ldots, t-1; \text{ and } t = 3, \ldots, T, \tag{11.4}$$
$$E(v_{i,t} \Delta y_{i,t-r}) = 0; \qquad E(v_{i,t} \Delta x_{i,t-r}) = 0;$$
$$\text{where } r = 1; \text{ and } t = 3, \ldots, T. \tag{11.5}$$

This estimator combines the $T-2$ equations in differences with the $T-2$ equations in levels into a single system. It uses the lagged levels of dependent and independent variables as instruments for the difference equation and the lagged differences of dependent and independent variables as instruments for the level equation. Blundell and Bond (1998) showed that this new system GMM estimator results in consistent and efficient parameter estimates, and has better asymptotic and finite sample properties.

We examined the nature of our data to determine whether the series for exporting and productivity are persistent. Our estimates of the AR (1) coefficients on exporting and productivity show that the series of exporting and productivity are highly persistent, thus the lagged levels of exports and productivity provide weak instruments for the differences in the first-differenced GMM model. As a result, we believe the system GMM estimator to be more appropriate than the first-differenced estimator in the context of this study.

Thus, we combine the first-differenced version of the ECM with the level version of the model, for which the instruments used must be orthogonal to the plant-specific effects. Note that the level of the dependent productivity variable must be correlated with the plant-specific effects, and we want to allow for the levels of the independent export variable to be potentially correlated with the plant specific effect. This rules out using the levels of any variables as instruments for the level equation. However, Blundell and Bond (1998) show that in autoregressive-distributed lag models, first differences of the series can be uncorrelated with the plant specific effect provided that the series have stationary means. In summary, the system GMM estimator uses lagged levels of the productivity and exports variables as instruments for the first-difference equation and lagged difference of the productivity and exports variables as instruments for the level form of the model.

This system GMM estimator results in consistent and efficient parameter estimates, and has good asymptotic and finite sample properties (relative to just straightforward estimation of first differences). Moreover, this estimation procedure allows us to examine the cross-sectional relationship between the levels of exporting and productivity since the firm-specific effect is not eliminated but rather controlled by the lagged differences of the dependent and independent variables as instruments, assuming that the differences are not correlated with a plant-specific effect, while levels are.

To determine whether our instruments are valid in the system GMM approach, we use the specification tests proposed by Arellano and Bond (1991) and Arellano and Bover (1995). First, we apply the Sargan test, a test of overidentifying restrictions, to determine any correlation between instruments and errors. For an instrument to be valid, there should be no correlation between the instrument and the error terms. The null hypothesis is that the instruments and the error terms are independent. Thus, failure to reject the null hypothesis could provide evidence that valid instruments are used. Second, we test whether there is a second-order serial correlation with the first differenced errors. The GMM estimator is consistent if there is no second-order serial correlation in the error term of the first-differenced equation. The null hypothesis in this case is that the errors are serially uncorrelated. Thus, failure to reject the null hypothesis could supply evidence that valid orthogonality conditions are used and the instruments are valid. One would expect the differenced error term to be first-order serially correlated, although the original error term is not. Finally, we use the differenced Sargan test to determine whether the extra instruments implemented in the level equations are valid. We compare the Sargan test statistic for the first-differenced estimator and the Sargan test statistic for the system estimator.

11.3.3. The data

This study uses unbalanced panel data on plants with more than 25 employees for the T&A industry (ISIC 3212 and 3222), and MV&P industry (ISIC 3843) industries from 1987–1997. Our sample represents a large fraction of the relevant population; textile (manufacture of textile goods except wearing apparel, ISIC 3212) and apparel (manufacture of wearing apparel except fur and leather, ISIC 3222) are subsectors of the textile, wearing apparel and leather industry (ISIC 32), which accounts for 35 percent of the total manufacturing employment, nearly 23 percent of wages, 20 percent of the output produced in the total manufacturing industry and approximately 48 percent of Turkish manufactured exports. The motor vehicles and parts industry (ISIC 3843) accounts for 5 percent of the total manufacturing employment, nearly 6.6 percent of wages, 10 percent of the output produced in the total manufacturing industry, and approximately 5.2 percent of Turkish manufactured exports. Thus, the data that is used in the study accounts for 53.2 percent of the total Turkish merchandise exports.

The data was collected by the State Institute of Statistics in Turkey from the Annual Surveys of Manufacturing Industries, and classified based on the International Standard Industrial Classification (ISIC Rev.2). These plant-level data consist of output; capital, labor, energy, and material inputs; investments; depreciation funds; import; export; and several plant characteristics.[5] These plant-level variables were also used to estimate the productivity indices (i.e. total factor productivity (TFP) and labor productivity (LP)) using the Multilateral Index approach of Good *et al.* (1996).[6]

[5] In the interest of space, a detailed description of how these variables are constructed is not presented here, but is available from the authors upon request.

[6] We acknowledge that there may be conceptual difficulties in the use of this type of productivity measure in our analysis (as raised by Katayama *et al.*, 2003). However, all

An important issue to note here is that some of the export values in our data set were missing for the years 1987, 1988, 1989 and 1997; even though the data for the other variables is complete. Instead of dropping these years, we chose to augment the export data by using interpolation and extrapolation techniques. To describe these techniques, consider a time series with n time periods: $x_1, x_2, \ldots, x_n$. Interpolation fills in missing values for all observations with missing data between non-missing observations. That is, if we have non-missing observations for $x_2, \ldots, x_8$ and $x_{11}, \ldots, x_{15}$, then interpolation will produce estimates for x_9 and x_{10}, but not for x_1. Extrapolation, on the other hand, fills in missing values both between and beyond non-missing observations. Thus, if the series ran from x_1 to x_{20}, then extrapolation will produce estimates for x_9 and x_{10}, but for x_1 and $x_{16}, \ldots, x_{20}$ as well. This approach for dealing with missing data is not new and has been used in other studies (Efron, 1994; Little and Rubin, 1987; Moore and Shelman, 2004; Rubin, 1976, 1987, 1996; Schafer, 1997; Schenker *et al.*, 2004). It is important to emphasize here that this augmentation did not markedly change the relevant results (i.e. signs, magnitude, and significance) of the estimated ECM model parameters using only the non-missing data (1990–96).[7]

Our models are estimated using the plants that export continuously, plants that begin as non-exporters during the first two years of the sample time period and become exporters thereafter and stayed in the market continuously, and the plants that start as exporters and exit during the last two years of the sample period. Plants that do not export at any point in the time period and the plants that enter and exit the export market multiple times are excluded.

In general, the nature of our panel data is such that the cross-section component is large but the time-series component is small. Hence, the system GMM estimation procedure above would be appropriate in this case. Summary statistics of all the relevant variables are presented in Table 11.1.

11.4. Results

The estimated parameters of various ECMs are presented in Tables 11.2 and 11.3, respectively. Specifically, Table 11.2 shows the estimated relationships between TFP growth and export growth, while Table 11.3 shows

previous studies in the literature still use this type of productivity measure and a feasible/refined alternative estimation procedure for an appropriate productivity measure has not been put forward. A more detailed discussion of the approach used for calculating the plant-level TFPs can be seen in Appendix A11.

[7] In the interest of space, the results for the non-augmented export data from 1990–96 are not reported here but are available from the authors upon request.

***Table 11.1.* Descriptive statistics from 1987–1997 (Y, K, E, and M are in Constant Value Quantities at 1987 Prices, in ‘000 Turkish Liras; L is in total hours worked in production per year)**

	Statistics			
	Mean	Standard deviation	Minimum	Maximum
A. Apparel and textile industries				
Output (Y)	1,951.53	4,386.20	3.09	115,994.70
Material (M)	1,690.16	3,428.79	0.02	90,651.77
Labor (L)	172.39	321.15	0.01	7,960.79
Energy (E)	48.85	316.51	0.02	13,469.06
Capital (K)	1,327.99	12,395.41	0.136	500,876.40
Small	0.585			
Medium	0.203			
Large	0.212			
TFP growth ($\Delta \ln \text{TFP}$)	0.011			
LP growth ($\Delta \ln \text{LP}$)	0.032			
Export growth ($\Delta \ln \text{EXP}$)	0.081			
Number of observations	7453			
Number of plants	1265			
B. MV&P industry				
Output (Y)	12,303.93	59,821.87	23.93	1,212,264.13
Material (M)	8,269.86	41,860.24	0.41	793,683.63
Labor (L)	336.62	933.14	0.01	18,141.64
Energy (E)	237.28	1,023.14	0.13	22,178.53
Capital (K)	5,506.36	31,480.68	0.60	720,275.88
Small[*]	0.514			
Medium	0.173			
Large	0.313			
TFP growth ($\Delta \ln \text{TFP}$)	0.060			
LP growth ($\Delta \ln \text{LP}$)	0.056			
Export growth ($\Delta \ln \text{EXP}$)	0.165			
Number of observations	2211			
Number of plants	328			

[*] We divide the plants into three size groups: small plants, with less than 50 employees; medium plants, between 50 and 100 employees; and large plants, with 100 employees or more.

the estimated relationships between LP growth and export growth. In these estimated models, we used total factor productivity (TFP) and labor productivity (LP) as our measures of productivity.

Note that in our discussion of Equation (11.2) in the previous section, we assume that the productivity measure is the dependent variable and amount of exports is one of the independent variables. This implies that the direction of causation is from exports to productivity. Although there are empirical studies that support the causal direction from exports to productivity (see, for example, Kraay, 1997; Bigsten *et al.*, 2002; Castellani, 2001), a number of empirical studies have also shown that the direction of causation may be in the other direction (see Bernard and Jensen, 1999; Aw *et al.*, 1998; Clerides *et al.*, 1998). Hence, aside from estimating the specification in Equation (11.2), where the productivity measure is the dependent variable, we also estimated ECM's where the amount of exports is the dependent variable and the productivity measure is the independent variable. In Tables 11.2 and 11.3, the first column for each industry shows the effect of the productivity measure (i.e. either TFP or LP) on exporting, while the second column for each industry shows the effect of exporting on the productivity measure.

As can be seen from these tables, the specification tests to check the validity of the instruments are satisfactory. The test results show no evidence of second-order serial correlation in the first differenced residuals. Moreover, the validity of lagged levels dated $t - 2$ and earlier as instruments in the first-differenced equations, combined with lagged first differences dated $t - 2$ as instruments in the levels equations are not rejected by the Sargan test of overidentifying restrictions.

The coefficients associated with the error correction terms in all the regression equations are significant and negative as expected. Thus, the results show that there is a strong long-run relationship between exporting and productivity. Furthermore, statistical significance of the error correction terms also imply that, when there are deviations from long-run equilibrium, short-run adjustments in the dependent variable will be made to re-establish the long-run equilibrium.

Now let us discuss each table in turn. In Table 11.2, for the equations where export growth is the dependent variable, the error correction coefficients have statistically significant, negative signs in both industries. However, the magnitude of the coefficients is different in each industry. This means that the speed of the short-run adjustment is different for the two industries. For the apparel and textile industries, the model converges quickly to equilibrium, with about 30 percent of discrepancy corrected in each period (coefficient of −0.302). The speed of the adjustment from the deviation in the long-run relationship between exports and productivity is slower in the MV&P industry (−0.114) relative to the T&A industry. On the other hand, for the equation where TFP growth is the dependent variable, the magnitudes of the error correction coefficients for the MV&P industry is greater than the T&A industry ($0.428 > 0.218$). This means that the speed of adjustment of TFP to temporary export shocks is slower in the T&A industry as compared to the MV&P industry.

Table 11.2. Estimated error correction model: long-run and short-run dynamics of TFP and exports (1987–1997)

Explanatory variables	Dependent variables			
	Apparel and textile industries		Motor vehicle and parts industry	
	$\Delta \ln \text{EXP}$	$\Delta \ln \text{TFP}$	$\Delta \ln \text{EXP}$	$\Delta \ln \text{TFP}$
$\Delta \ln \text{TFP}$	1.099		0.567	
	(0.285)*		(0.212)*	
$\Delta \ln \text{TFP}_{t-1}$	−0.494	0.358	−0.181	0.492
	(0.152)*	(0.046)*	(0.107)***	(0.025)*
$\ln \text{TFP}_{t-2}$	−0.022		−0.070	
	(0.102)		(0.077)	
$\ln \text{TFP}_{t-2} - \ln \text{EXP}_{t-2}$		−0.218		−0.428
		(0.030)*		(0.040)*
$\Delta \ln \text{EXP}$		0.160		0.035
		(0.051)*		(0.039)
$\Delta \ln \text{EXP}_{t-1}$	0.626	−0.080	0.287	−0.026
	(0.124)*	(0.035)**	(0.047)*	(0.021)
$\ln \text{EXP}_{t-2}$		−0.192		−0.410
		(0.032)*		(0.048)*
$\ln \text{EXP}_{t-2} - \ln \text{TFP}_{t-2}$	−0.302		−0.114	
	(0.066)*		(0.028)*	
Summation of short-run coef.	0.605	0.080	0.386	0.009
Short-run wald test (P-value)	0.001	0.007	0.026	0.480
Long run coefficient	0.927	0.119	0.391	0.043
Long run coefficient (P-value)	0.000	0.000	0.013	0.000
Sargan difference test (P-value)	0.267	0.281	0.144	0.776
Sargan test (P-value)	0.333	0.311	0.277	0.363
AR1 (P-value)	0.306	0.000	0.020	0.037
AR2 (P-value)	0.107	0.675	0.698	0.110
Number of observations	3778		932	
Number of plants	661		116	

Notes: (1) Estimation by System-GMM using DPD for OX (Doornik *et al.*, 2002). (2) Asymptotically robust standard errors are reported in parentheses. (3) The Sargan test is a Sargan–Hansen test of overidentifying restrictions. The null hypothesis states that the instruments used are not correlated with the residuals. (4) AR1 and AR2 are tests for first- and second-order serial correlation in the first-differenced residuals. The null hypothesis for the second-order serial correlation test states that the errors in the first-differenced regression do not show second-order serial correlation. (5) Lagged levels of productivity and exports (dated $t-2$ and earlier) in the first-differenced equations, combined with lagged first differences of productivity and exports (dated $t-2$) in the level equations are used as instruments. (6) Year dummies are included in each model.

*Significant at the 1% level.

**Significant at the 5% level.

***Significant at the 10% level.

Table 11.3. Estimated error correction model: long-run and short-run dynamics of lp and exports (1987–1997)

Explanatory variables	Dependent variables			
	Apparel and textile industries		Motor vehicle and parts industry	
	$\Delta \ln \mathrm{EXP}$	$\Delta \ln \mathrm{TFP}$	$\Delta \ln \mathrm{EXP}$	$\Delta \ln \mathrm{TFP}$
$\Delta \ln \mathrm{LP}$	0.857		0.429	
	(0.131)*		(0.195)**	
$\Delta \ln \mathrm{LP}_{t-1}$	−0.435	0.320	−0.148	0.345
	(0.060)*	(0.068)*	(0.089)	(0.037)*
$\ln \mathrm{LP}_{t-2}$	0.069		−0.026	
	(0.042)***		(0.057)	
$\ln \mathrm{LP}_{t-2} - \ln \mathrm{EXP}_{t-2}$		−0.157		−0.240
		(0.025)*		(0.051)*
$\Delta \ln \mathrm{EXP}$		0.231		0.182
		(0.080)*		(0.071)*
$\Delta \ln \mathrm{EXP}_{t-1}$	0.432	−0.108	0.316	−0.074
	(0.045)*	(0.067)	(0.049)*	(0.037)**
$\ln \mathrm{EXP}_{t-2}$		−0.107		−0.159
		(0.017)*		(0.045)*
$\ln \mathrm{EXP}_{t-2} - \ln \mathrm{LP}_{t-2}$	−0.377		−0.135	
	(0.037)*		(0.024)*	
Summation of short-run coef.	0.422	0.123	0.281	0.108
Short-run wald test (P-value)	0.000	0.016	0.080	0.036
Long run coefficient	1.183	0.319	0.807	0.338
Long run coefficient (P-value)	0.000	0.000	0.000	0.000
Sargan difference test (P-value)	0.210	0.126	0.918	0.353
Sargan test (P-value)	0.208	0.296	0.707	0.677
AR1 (P-value)	0.046	0.023	0.032	0.004
AR2 (P-value)	0.905	0.790	0.548	0.329
Number of observations	3778		932	
Number of plants	661		116	

Notes: (1) Estimation by system-GMM using DPD for OX (Doornik *et al.*, 2002). (2) Asymptotically robust standard errors are reported in parentheses. (3) The Sargan test is a Sargan–Hansen test of overidentifying restrictions. The null hypothesis states that the instruments used are not correlated with the residuals. (4) AR1 and AR2 are tests for first- and second-order serial correlation in the first-differenced residuals. The null hypothesis for the second-order serial correlation test states that the errors in the first-differenced regression do not show second-order serial correlation. (5) Lagged levels of productivity and exports (dated $t-2$ and earlier) in the first-differenced equations, combined with lagged first differences of productivity and exports (dated $t-2$) in the level equations are used as instruments. (6) Year dummies are included in each model.

*Significant at the 1% level.

**Significant at the 5% level.

***Significant at the 10% level.

Another thing to note in Table 11.2 is the different short-run behaviors in the T&A versus the MV&P industry. In the T&A industry, the speed of short-run export adjustment as a response to temporary TFP shocks (−0.302) tend to be faster than the short-run TFP adjustments to temporary export shocks (−0.218). In contrast, for the MV&P industry, the speed of short-run export adjustment as a response to temporary TFP shocks (−0.114) tend to be slower than the short-run TFP adjustments to temporary export shocks (−0.428). This suggests that short-run industrial policy may need to be treated differently in both these industries due to the dissimilar short-run dynamics.

The pattern of results in Table 11.3 is the same as the ones in Table 11.2. That is, the speed of short-run export adjustments to LP shocks tend to be faster in the T&A industry (−0.377) as compared to the MV&P industry (−0.135). In contrast, the speed of adjustment of LP to temporary export shocks is slower in the T&A industry (−0.157) as compared to the MV&P industry (−0.240). Also, the speed of short-run export adjustment as a response to temporary LP shocks (−0.377) tend to be faster than the short-run LP adjustments to temporary export shocks (−0.157). Conversely, for the MV&P industry, the speed of short-run export adjustment as a response to temporary LP shocks (−0.135) tend to be slower than the short-run TFP adjustments to temporary export shocks (−0.24). These results again suggest that the potential industry-specific short-run impacts should be taken into account setting temporary industrial policy.

The coefficient of the error correction term gives us an indication of the speed of adjustment, but it is also important to examine the magnitudes of the short-run effects as measured by the short-run coefficient. From Equation (11.2), the short-run coefficient is computed by adding the coefficients of the contemporaneous and lagged dependent variable. From Tables 11.2 and 11.3, it is evident that the magnitude of the short-run export response to a temporary productivity shock is greater than the short-run productivity effect of a temporary export shock (in both industries). This suggests that temporary shocks in productivity will result in bigger short-run export adjustments relative to the converse.

Aside from short-run adjustments of the variables, it is also important to examine the long-run relationships implied from the ECMs. For this we use the long-run elasticities of the dependent variables to the independent variables (see Tables 11.2 and 11.3). These long-run elasticities are calculated by subtracting the ratio of the coefficient of the scale effect (lag value of independent variable) to the coefficient of the error correction term from one. The statistical significance of these elasticities is tested with a Wald test. The test results indicate that the estimated long-run elasticities for all the estimated equations are statistically significant (at the 5% level) in both industries.

In Table 11.2, for the equations where export growth is the dependent variable, the long-run elasticities indicate that long-run export response to permanent shocks in TFP is large (for both industries). On the other hand, for the case where TFP is the dependent variable, the long-run elasticities suggest that long-run TFP adjustments to permanent changes in exports are lower. The results are very similar for the case of LP (Table 11.3). The long-run elasticities reveal that long-run export response to permanent shocks in LP is tend to be greater than the LP response to permanent changes in exports (for both industries).

Overall, the results from the long-run elasticities show that productivity response to permanent shocks in exports is lower than the export response to the permanent shocks in productivity for both the T&A and MV&P industries. Moreover, our analysis of short-run dynamics reveals that, for the MV&P industry, short-run productivity adjustments to temporary shocks in exports tend be faster than the short-run export adjustments to temporary shocks in productivity. For the apparel industry, short-run export adjustments due to temporary productivity shocks are faster relative to the short-run productivity adjustments from temporary export shocks. However, the estimated short-run coefficient in both industries indicates that short-run productivity response to temporary export shocks is larger than the short-term export response to temporary productivity shocks. These results suggest similar behaviors in terms of the magnitudes of the short-run and long-run effects of exports/productivity shocks. But speed of short-run adjustments tends to be different depending on the type of industry. Knowledge of these plant behaviors can help improve the design of industrial policies that would allow further economic growth in Turkey.

11.5. Conclusions and policy implications

In this paper, we examine the short-run and the long-run dynamics of the relationship between export levels and productivity for two Turkish manufacturing industries. An error correction model is estimated using a system GMM estimator to overcome problems associated with unobserved plant-specific effects, persistence, and endogeneity. This approach allows us to obtain consistent and efficient estimates of the short-run and long-run relationships of exports and productivity. From these estimates, we conclude that permanent productivity shocks induce larger long-run export level responses, as compared to the effect of permanent export shocks on long-run productivity. A similar behavior is evident with respect to the magnitude of the effects of temporary shocks on short-run behavior. In addition, for the T&A industry, our short-run analysis shows that temporary export shocks usually result in faster short-run productivity adjustments, as compared to the effects of productivity shocks on short-run exports. The converse is true for the MV&P industry.

From an industrial policy perspective, our analysis suggests that policies which induce permanent productivity enhancements would result in large long-run export effects. Hence, policies aimed at permanently improving productivity should be implemented by the policy makers to obtain sustainable export performance and a bigger role in the global market. This may then lead to more sustained economic growth. This insight may help explain the apparent failure of the trade liberalization policies in the 1980s to sustain productivity and growth in the economy. Most developing countries enact policies to promote exports on the assumption that it will be good for productivity and economic growth. From our results, there would be a positive productivity response if this was a permanent promotion policy, but this kind of policy would still only generate a small long-term effect on productivity and/or economic growth.

In addition, if the export promotion policy is temporary, there would probably be differential short-run speed of adjustments depending on the type of industry where it is implemented. For the MV&P industry, our results suggest that a temporary export promotion policy would result in a fast productivity response. But in the T&A industry, a temporary export promotion policy may lead to a slower productivity response. The reason is that the MV&P industry in Turkey tends to constitute large plants that heavily invest in technology, while plants in the apparel industry tend to be small to medium sized with less investments in technology. Hence, if government policy makers want to implement short-run policies to show fast performance effects, then they must consider the short-run dynamic behavior of plants in different industries in their decision-making. On the other hand, in both the T&A and MV&P industry, there would be larger short-run export adjustments from temporary shocks in productivity relative to the short-run productivity enhancements from temporary export shocks. This is consistent with our long-run insights that productivity enhancements tend to have larger export effects, which again point to the appropriateness of enacting productivity-enhancing policies as the main tool for driving export and economic growth.

Acknowledgements

We are greatly indebted to Omer Gebizlioglu, Ilhami Mintemur and Emine Kocberber at State Institute of Statistics in Turkey for allowing us to have an access to the data for this study. We also thank discussants

and participants at the 11th International Conference on Panel Data, 2004 North American Summer Meeting of the Econometric Society, 2004 International Industrial Organization Conference for their suggestions that helped us greatly in revising the paper.

Appendix A11. Calculation of plant-level total factor productivity

The main measure of productivity used in this study is total factor productivity (TFP). In the plant-level analysis, we construct a multilateral index to measure the plant-level TFP for the period 1987–1997. In this study, we use Good *et al.* (1996) approach for computing the multilateral TFP index. In their approach, different hypothetical plant reference points are constructed for each cross-section, and then the hypothetical plants are linked together over time. This type of multilateral index has the advantage of providing measures either from year to year or from a sequence of years, through the process of chain-linking.

In this study, the multilateral TFP index measure for plant j, which produces a single output Y_{jt} using inputs X_{ijt} with cost shares S_{ijt}, is calculated as follows:

$$\begin{aligned}\ln \mathrm{TFP}_{jt} = {} & (\ln Y_{jt} - \overline{\ln Y_t}) + \sum_{k=2}^{t}(\overline{\ln Y_k} - \overline{\ln Y}_{k-1}) \\ & - \left[\sum_{i=1}^{n}\frac{1}{2}(S_{ijt} + \bar{S}_{it})(\ln X_{ijt} - \overline{\ln X}_{it}) \right. \\ & \left. + \sum_{k=2}^{t}\sum_{i=1}^{n}\frac{1}{2}(\bar{S}_{ik} + \bar{S}_{ik-1})(\overline{\ln X}_{ik} - \overline{\ln X}_{ik-1})\right] \qquad \text{(A11.1)}\end{aligned}$$

where $\overline{\ln Y_t}$ and $\overline{\ln X}_{it}$ are the natural log of the geometric mean of output and the natural log of the geometric mean of the inputs (capital, energy, labor, and material inputs) across all plants in time t, respectively. The subscript j represents individual plants such that $j = 1, 2, \ldots, N$. The subscript i is used to represent the different inputs where $i = 1, 2, \ldots, n$. The subscript k represents time period from $k = 2, 3, \ldots, t$ (i.e. if we are considering 10 years in the analysis, $k = 2, 3, \ldots, 10$).

The first two terms in the first line measure the plant's output relative to the hypothetical plant in the base year. The first term describes the deviation between the output of plant j and the representative plant's output, $\overline{\ln Y_t}$, in year t. This first sum allows us to make comparisons between cross-sections. The second term sums the change in the hypothetical plant's output across all years, while chaining the hypothetical plant values back to the base year. This allows us to measure the change in output of a typical plant over years. The following terms provide similar information. However, it is for inputs using cost shares and arithmetic average cost shares in each year as weights. Cost shares are just the proportion of the cost of input i relative to the total cost of all inputs. The resulting measure is the total factor productivity of plant j in year t relative to the hypothetical plant in the base year (1987, in this case). With this measure the distribution of plant-level total factor productivity can then be analyzed.

Aside from the plant-level TFP, we also calculate the plant-level labor productivity using the same multilateral index calculation described above but only using labor on the input side of the calculation. Using the labor productivity in our analysis ensures that our analysis is robust to any changes in productivity measure used. However, it is important to note that labor productivity is only a partial measure of TFP and has its own shortcomings. For example, if the production technology among plants within the industry differs such that they do not have similar input–output ratios, then labor productivity is not a good measure of efficiency and may be a misleading measure of performance. The TFP may be more appropriate in this case. Nevertheless, using the labor productivity would allow us to somehow assess the robustness of our results.

We use kernel density estimates for the plant-level TFP and labor productivity measure to summarize the distribution of plant productivity. Figures A11.1–A11.4 show kernel density estimates of TFP and labor productivity for the industries, respectively. The kernel density was estimated for two time periods, from 1987–1993 and 1994–1997. These two time periods were chosen because the country suffered an economic crisis in 1994 and soon afterwards the government introduced institutional changes in their economic and financial policies. For example, the Economic Stabilization and Structural Adjustment Program was enacted, where export-oriented policies such as subsidies and wage suppression were put in place. The foreign exchange system was regulated and the capital inflow was controlled during this period. Prior to 1994, the pre-dominant policies of the government were the opposite of the post-1994 period (i.e. the government relinquished control of capital markets, eliminate subsidies, and increase wages). Hence, the period 1987–93 can be called the pre-crisis period and the period 1994–1997 is the post-crisis period.

Figures A11.3 and A11.4 clearly show that there is a slight rightward shift in TFP and labor productivity during the post-crisis period in Motor vehicle and Parts industry. For the Apparel and Textile industry there is a rightward shift in the labor productivity but not in the TFP.

Figure A11.1. Kernel density estimate of the distribution of total factor productivity in the apparel and textile industry

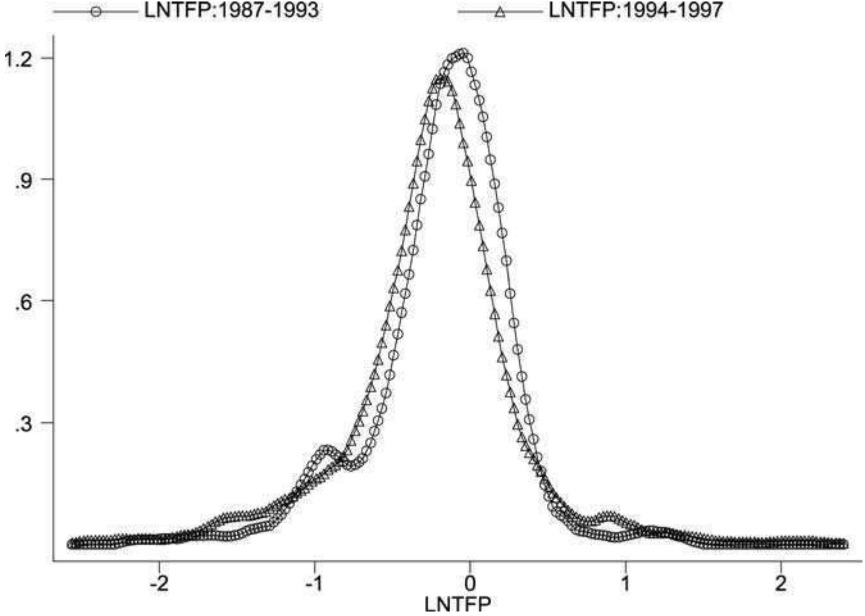


Figure A11.2. Kernel density estimate of the distribution of labor productivity in the apparel and textile industry

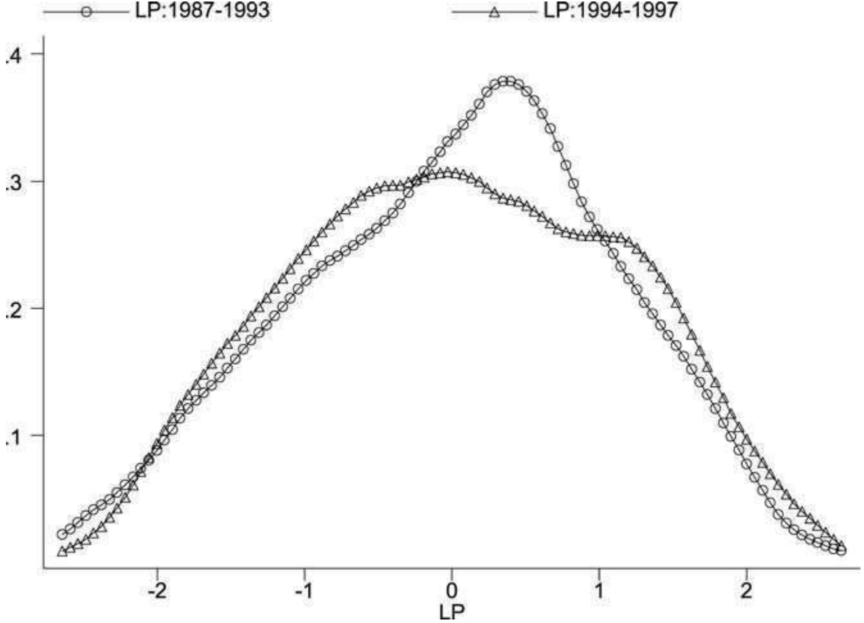


Figure A11.3. Kernel density estimate of the distribution of total factor productivity in the motor vehicle and parts industry

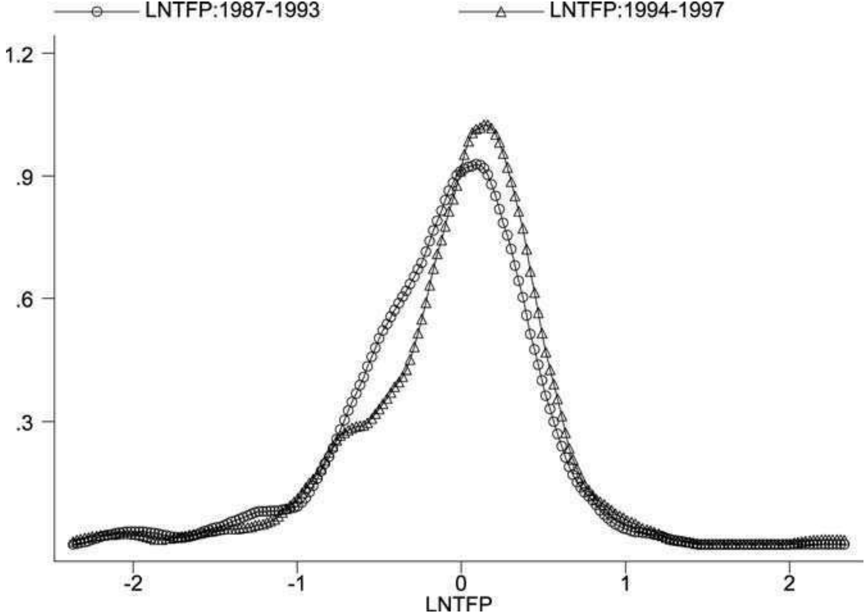
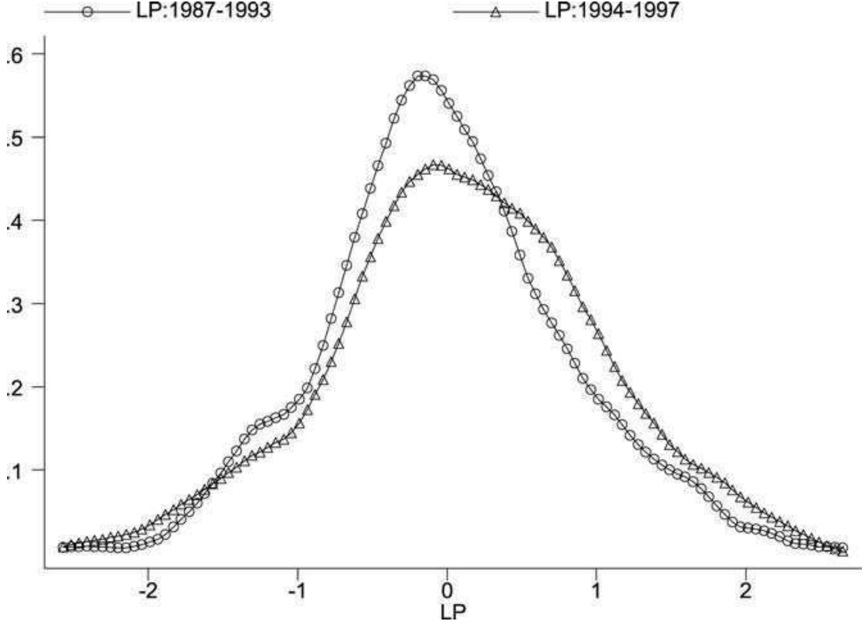


Figure A11.4. Kernel density estimate of the distribution of labor productivity in the motor vehicle and parts industry



Appendix B11. Plant performance of exporters and non-exporters: export premia

This section reports estimates of the proportional differences between the characteristics of exporting and non-exporting plants in the Turkish apparel and textile and motor vehicle and parts industries by forming the

Table B11.1. Plant performance of exporters and non-exporters: export premia

Dependent variables	Export premia in the apparel and textile industries	Export premia in the motor vehicle and parts industry
Total factor productivity	0.082 (0.016)*	0.064 (0.031)*
Labor productivity	0.806 (0.025)*	0.368 (0.045)*
Wage per employee	0.126 (0.011)*	0.250 (0.031)*
Output per employee	0.796 (0.023)*	0.331 (0.042)*
Capital per employee	0.589 (0.045)*	0.458 (0.086)*
Capital in machine per employee	0.464 (0.046)*	0.599 (0.093)*
Imported capital stock per employee	0.418 (0.112)*	0.452 (0.146)*
Total investment per employee	0.452 (0.058)*	0.238 (0.107)**
Administrative labor	1.070 (0.025)*	1.678 (0.061)*
Labor hours	0.777 (0.021)*	1.269 (0.050)*
Total employment	0.794 (0.020)*	1.331 (0.051)*
Output	1.655 (0.031)*	1.969 (0.072)*

Notes: (1) Robust t-statistics are in parentheses. (2) The independent variables for the different regressions include time, size, and region dummies (except for the regression where total output, administrative labor, labor hours, and total employment were the dependent variables – these regressions do not include the size dummies). Dependent variables are in natural logs. The base group is non-exporters.

*Significant at the 1% level.

**Significant at the 5% level.

following regression (see Bernard and Wagner, 1997):

$$X_{it} = \alpha_0 + \alpha_1 \text{Exporter}_{it} + \alpha_2 \text{Size}_{it} + \alpha_3 \text{Region}_{it} + \alpha_4 \text{Year}_t + e_{it}, \tag{B11.1}$$

where X stands for either the log or the share of plant characteristics that reflect plant capabilities in productivity, technology, and employment.[8] Exporter is a dummy variable for the export status, taking a value of 1 if the plant exports in the current year. Year dummies are included to capture macroeconomic shocks and the changes in the institutional environment. The agglomeration effect might be important in explaining the differences in plant characteristics (see Krugman, 1991; Porter, 1998). There are large development disparities across Turkey's regions because of different regional capabilities such as infrastructure, rule of law, quality of public services, localized spillovers, the export and import density, foreign investment intensity (to take advantage of international spillovers, see Coe and Helpman, 1995). Therefore, we included regional dummies to correct for the exogenous disparities in the productivity differences across the regions. Finally, the plant size is included to capture differences in the production technology across plants of different sizes. One would expect the larger plants to be more productive for two reasons. They benefit from scale economies and have access to more productive technology to a greater extent. However, they tend to be less flexible in their operation which affects productivity negatively. In order to capture the size effects, we divide the plants into three size groups: small plants, with less than 50 employees; medium plants, with between 50 and 100 employees; and large plants, with 100 employees or more. We select the small group as the base group. The omitted variable is non-exporters. The coefficient on the exporting dummy variable, α_1, shows the average percentage difference between exporters and non-exporters, conditional on size, region, and year.

The estimated parameters are presented in Table B11.1. All of the estimated premia are statistically significant and positive. Our results show that the difference in total factor productivity between exporters and non-exporters is large and statistically significant for both industries. The exporting plants have significantly higher productivity for both industries. After controlling for region, year, and size, the difference in total factor productivity between exporting and non-exporting plants was highest in

[8] We also included the export intensity (the ratio of exports to output of the plant) as an explanatory variable in the regression; however, the results did not change. The export premia was greater for the plants that export a higher proportion of their output.

the apparel and textile industries at 8.2 percent, followed by the motor vehicle and parts industry at 6.4 percent. The difference in labor productivity between exporting and non-exporting plants was even more dramatic. The difference was positive and significant for each of the two industries, with the apparel industry at 80.6 percent and the motor vehicle industry at 36.8 percent.

Our results also show that exporting plants have significantly higher output. The exporters produce significantly more output per employee; the apparel and textile industries is 79.6 percent higher and the motor vehicle parts industry is 33.1 percent higher. Exporting plants are more capital-intensive, and they invest more heavily in machinery and equipment. Exporters also pay their workers, on average, significantly higher wages than non-exporters. Exporters in the apparel and textile and motor vehicle and parts industries pay 12.6 and 25 percent higher wages, respectively. In short, our results show that exporting plants perform much better than their domestically oriented counterparts.

References

Anderson, T.W., Hsiao, C. (1981), “Estimation of dynamic models with error components”, *Journal of the American Statistical Association,* Vol. 76, pp. 598–606.

Arellano, M., Bond, S.R. (1991), “Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations”, *Review of Economic Studies*, Vol. 58, pp. 277–297.

Arellano, M., Bover, O. (1995), “Another look at the instrumental variable estimation of error components models”, *Journal of Econometrics,* Vol. 68, pp. 29–52.

Aw, B., Chung, S., Roberts, M. (1998), *Productivity and the Decision to Export Market: Micro Evidence from Taiwan and South Korea*, NBER, Cambridge, MA.

Baltagi, B.H. (2001), *Econometric Analysis of Panel Data*, 2nd ed., Wiley, New York.

Banerjee, A., Galbraith, J., Dolado, J.J. (1990), “Dynamic specification with the general error-correction form”, *Oxford Bulletin of Economics and Statistics*, Vol. 52, pp. 95–104.

Banerjee, A., Dolado, J.J., Galbraith, J., Hendry, D.F. (1993), *Cointegration, Error Correction, and the Econometric Analysis of Non-stationary Data,* Oxford University Press, Oxford.

Banerjee, A., Dolado, J.J., Mestre, R. (1998), “Error-correction mechanisms tests for cointegration in a single-equation framework”, *Journal of Time Series Analysis,* Vol. 19, pp. 267–284.

Barro, R.J., Sala-I-Martin, X. (1995), *Economic Growth,* McGraw-Hill, New York.

Bernard, A.B., Jensen, J.B. (1998), “Exporters, jobs and wages in U.S. manufacturing, 1976–1987”, The Brooking Papers on Economic Activity, pp. 67–112.

Bernard, A.B., Jensen, J.B. (1999), “Exceptional exporter performance: cause, effect, or both”, *Journal of International Economics,* Vol. 47, pp. 1–26.

Bernard, A.B., Jensen, J.B. (2001), “Why some firms export”, NBER Working Paper No. 8349, NBER.

Bernard, A.B., Wagner, J. (1997), "Exports and success in German manufacturing", *Weltwirtschaftliches Archive*, Vol. 133, pp. 134–157.

Bigsten, A., Collier, P., *et al.* (2002), "Do African manufacturing firms learn from exporting?", Centre for the Study of African Economies Working Paper Series, WPS/2002-09, Oxford University, Oxford.

Binder, M., Hsiao, C., Pesaran, M.H. (2005), "Estimation and inference in short panel vector autoregressions with unit roots and contegration", *Econometric Theory*, Vol. 21 (4), pp. 795–837.

Blundell, R., Bond, S.R. (1998), "Initial conditions and moment restrictions in dynamic panel data models", *Journal of Econometrics*, Vol. 87, pp. 115–143.

Blundell, R., Bond, S.R. (2000), "GMM estimation with persistent panel data: an application to production functions", *Econometric Reviews*, Vol. 19, pp. 321–340.

Blundell, R., Bond, S.R., Windmeijer, F. (2000), "Estimation in dynamic panel data models: improving on the performance of the standard GMM estimators", The Institute of Fiscal Studies, London.

Bond, S.R., Elston, J., Mairesse, J., Mulkay, B. (1997), "Financial factors and investment in Belgium, France, Germany and the UK: a comparison using company panel data", NBER Working Paper, NBER, Cambridge, MA.

Bond, S.R., Harhoff, D., Reenen, J.V. (1999), "Investment, R&D, and financial constraints in Britain and Germany", Mimeo, Institute for Fiscal Studies, London.

Castellani, D. (2001), "Export behavior and productivity growth: evidence from Italian manufacturing firms", Mimeo, ISE-Università di Urbino.

Chao, W.S., Buongiorno, J. (2002), "Exports and growth: a causality analysis for the pulp and paper industries based on international panel data", *Applied Economics*, Vol. 34, pp. 1–13.

Clerides, S.K., Lach, S., Tybout, J.R. (1998), "Is learning-by-exporting important? Micro dynamic evidence from Colombia, Mexico, and Morocco", *Quarterly Journal of Economics*, Vol. 113, pp. 903–947.

Coe, D., Helpman, E. (1995), "International R&D spillovers", *European Economic Review*, Vol. 39, pp. 859–887.

Davidson, J.E.H., Hendry, D.F., Srba, F., Yeo, S. (1978), "Econometric modeling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom", *Economic Journal*, Vol. 88, pp. 661–692.

Dodaro, S. (1991), "Comparative advantage, trade, and growth: export-led growth revisited", *World Development*, Vol. 19, pp. 1153–1165.

Doornik, J., Arellano, M., Bond, S. (2002), "Panel data estimation using DPD for OX", Nuffield College, Oxford.

Efron, B. (1994), "Missing data, imputation, and the bootstrap", *Journal of the American Statistical Association*, Vol. 89, pp. 463–478.

Girma, S., Greenaway, D., Kneller, R. (2003), "Export market exit and performance dynamics: a causality analysis of matched firms", *Economics Letters*, Vol. 80, pp. 181–187.

Good, D., Nadiri, I., Sickles, R. (1996), "Index number and factor demand approaches to the estimation of productivity", NBER Working Paper No. 5790, NBER.

Grossman, G., Helpman, E. (1991), *Innovation and Growth in the Global Economy*, MIT Press, Cambridge.

Im, K.S., Pesaran, M.H., Shin, Y. (1997), "Testing for unit roots in heterogeneous panels", Mimeo, Cambridge, available on www.econ.cam.ac.uk/faculty/pesaran.

Katayama, H., Lu, S., Tybout, J. (2003), "Why plant-level productivity studies are often misleading and an alternative approach to inference", NBER Working Paper No. 9617, NBER.

Kraay, A. (1997), "Exports and economic performance: evidence from a panel of Chinese enterprises", Working Paper, World Bank, Washington, DC.

Krugman, P. (1991), "Increasing returns and economic geography", *Journal of Political Economy,* Vol. 99, pp. 483–499.

Levin, A., Lin, C. (1993), "Unit root tests in panel data: new results", U.C. San Diego Working Paper.

Little, R.J.A., Rubin, D.B. (1987), *Statistical Analysis with Missing Data,* Wiley, New York.

Lucas, R.E. (1988), "On the mechanics of economic development planning", *Journal of Monetary Economics,* Vol. 22, pp. 3–42.

Mairesse, J., Hall, B.H., Mulkay, B. (1999), "Firm-level investment in France and the United States: an exploration of what we have learned in twenty years", *Annales d'Economie et de Statistiques,* Vol. 55–56, pp. 27–67.

Marschak, J., Andrews, W. (1944), "Random simultaneous equations and the theory of production", *Econometrica,* Vol. 12, pp. 143–205.

Moore, W.H., Shelman, S.M. (2004), "Whither will they go? A global analysis of refugee flows, 1955–1995", Paper presented at the 2004 Annual Meeting of the Midwest Political Science Assoc. (April 15–18, 2004) Chicago, IL.

Parente, S., Prescott, E. (1994), "Barriers to technology adaptation and development", *Journal of Political Economy,* Vol. 102, pp. 298–321.

Porter, M.E. (1998), "Clusters and new economies of competition", Harvard Business Review, Nov.–Dec., pp. 7–90.

Pugel, T. (2004), *International Economics*, 12th ed., McGraw Hill, New York.

Rivera-Batiz, L.A., Romer, P. (1991), "Economic integration and endogenous growth", *Quarterly Journal of Economics,* Vol. 106, pp. 531–555.

Roberts, M., Tybout, J.R. (1997), "The decision to export in Colombia: an empirical model of entry with sunk costs", *American Economic Review,* Vol. 87, pp. 545–564.

Rubin, D.B. (1976), "Inference and missing data", *Biometrika,* Vol. 63, pp. 581–592.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Rubin, D.B. (1996), "Multiple imputation after 18+ years", *Journal of the American Statistical Association,* Vol. 91, pp. 473–489.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall, New York.

Schenker, N., Raghunathan, T., *et al.* (2004), "Multiple imputation of family income and personal earnings in the National Helath Interview Survey: methods and examples", Technical Document, National Center for Health Statistics, Center for Disease Control (CDC), Atlanta.

Wagner, J. (2002), "The causal effects of exports on firm size and labor productivity: first evidence from a matching approach", *Economics Letters,* Vol. 77, pp. 287–292.

This page intentionally left blank

Panel Data Econometrics
B.H. Baltagi (Editor)
© 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S0573-8555(06)74012-3

CHAPTER 12

Learning about the Long-Run Determinants of Real Exchange Rates for Developing Countries: A Panel Data Investigation

Imed Drine[a] and Christophe Rault[b]

[a]Paris I, Maison des Sciences de l'Economie, 106-112 bd. de L'Hôpital, 75647 Paris cedex 13, France
E-mail address: drine@univ-paris1.fr
[b]University of Evry-Val d'Essonne, Département d'économie, Boulevard François Mitterrand, 91025 Évry cedex, France
E-mail address: chrault@hotmail.com
url: http://www.multimania.com/chrault/index.html

Abstract

The main goal of this paper is to tackle the empirical issues of the real exchange rate literature by applying recently developed panel cointegration techniques developed by Pedroni ("Critical values for cointegrating tests in heterogeneous panels with multiple regressors", Oxford Bulletin of Economics and Statistics, Vol. 61 (Supplement) (1999), pp. 653–670; "Panel cointegration; asymptotic and finite sample properties of pooled time series tests with an application to the purchasing power parity hypothesis", Econometric Theory, Vol. 20 (2004), pp. 597–625) and generalized by Banerjee and Carrion-i-Silvestre ("Breaking panel data cointegration", Preliminary draft, October 2004, downloadable at http://www.cass.city.ac.uk/conferences/cfl/CFLPapersFinal/Banerjee%20and%20Carrion-i-Silvestre%202004.pdf) to a structural long-run real exchange rate equation. We consider here a sample of 45 developing countries, divided into three groups according to geographical criteria: Africa, Latin America and Asia. Our investigations show that the degrees of development and openness of the economy strongly influence the real exchange rate.

Keywords: real exchange rate, developing country

JEL classifications: E31, F0, F31, C15

12.1. Introduction

The relationship between the real exchange rate and economic development is certainly an important issue, both from the positive (descriptive) and normative (policy prescription) perspectives. In recent years, policy discussions have included increasing references to real exchange rate stability and correct exchange rate alignment as crucial elements to improve economic performance in emergent countries. Real exchange rate misalignment affects economic activity in developing countries mainly due to the dependence on imported capital goods and specialization in commodity exports. Accessibility to world financial markets which helps to smooth out consumption by financing trade imbalance, also plays an important role. Evidence from developing countries is often quoted to support the view that the link between real exchange rate misalignment and economic performance is strong. Cottani *et al.* (1990) argued that in many emergent countries, persistently misaligned exchange rate harmed the development of agriculture, reducing domestic food supply. Besides, a number of researchers have also pointed out the importance of understanding the main determinants of real exchange rate.

Edwards (1989) for instance developed a theoretical model of real exchange rate and provided an estimation of its equilibrium value for a panel of developing countries. According to these estimations, the most important variables affecting the real exchange rate equilibrium level are the terms of trade, the level and the composition of public spending, capital movements, the control of exchange and the movements of goods, technical progress, and capital accumulation.

Following Edwards's pioneering works applied studies estimating equilibrium exchange rates have increased these last past years, both for developed and developing countries. Among the large number of papers available in the literature, special attention should be drawn to the work of Xiaopu (2002), and Mac Donald and Ricci (2003) who investigated a number of issues that are relevant to an appropriate assessment of the real exchange rate equilibrium level and to the interesting review of literature for developing countries by Edwards, *NBER Working Papers*, 1999. In these studies the main long-run determinants of the real exchange rate are the terms of trade, the openness degree of the economy, and capital flows.

The aim of this paper is to apply recent advances in the econometrics of non-stationary panel methods to examine the main long-run determinants of the real exchange rate. We consider a sample of 45 developing countries, divided into three groups according to geographical criteria: *Africa* (21 countries: Algeria, Benin, Burkina Faso, Burundi, Cameroon, Congo, the democratic Republic of Congo, Ivory Coast, Egypt, Ethiopia, Gabon, Gambia, Ghana, Guinea Bissau, Kenya, Mali, Morocco, Mozambique, Niger, Senegal, Tunisia), *Latin America* (17 countries: Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, the Dominican Republic, Ecuador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Peru, Paraguay, Uruguay, Venezuela) and *Asia* (7 countries: Bangladesh, Indonesia, South Korea, India, Malaysia, the Philippines, Thailand). This grouping of countries according to a geographic criterion is justified by the fact that the panel data econometric techniques require a certain degree of homogeneity to get robust empirical results. This geographic criterion seems to us the most adapted and the most straightforward here for our sample of 45 countries especially since we wish to study the determinants of the real exchange rate for various continents. A grouping of countries according to an economic criterion would also have been possible but more complex to implement given the multiplicity of potential economic criteria.

The point here is to go beyond the teachings of the Balassa–Samuelson's theory (cf. in particular Drine and Rault, 2002; Drine *et al*., 2003 for these countries as well as Strauss, 1999 for OECD countries) and to determine if other factors, such as demand factors, economic policy or capital movements, also have an influence on the equilibrium real exchange rate level determination. Our econometric methodology rests upon the panel data integration tests proposed by Im *et al*. (1997, 2003) that assumes cross-sectional independence among panel units, Choi (2002) and Moon and Perron's (2003) (these two tests relaxing the assumption of cross-sectional independence which is often at odds with economic theory and empirical results), and on the panel data cointegration tests developed by Pedroni (1999, 2004) and generalized by Banerjee and Carrion-i-Silvestre (2004). The advantage of panel data integration and cointegration techniques is threefold: firstly, they enable to by-pass the difficulty related to short spanned time series, then they are more powerful than the conventional tests for time series and finally inter-individual information reduces the probability to make a spurious regression Banerjee (1999). To our best knowledge no comparable studies exist using these new econometric techniques to investigate the main macroeconomic variables influencing the real exchange rate in the long run in developing countries.

The remainder of the paper is organized as follows. In the second section we present a simple theoretical model of real exchange rate determination. In the third one we report and comment on our econometric results for a panel of 45 developing countries. A final section reviews the main findings. We find in particular, that besides the Balassa–Samuelson effect, other macroeconomic variables, such as the terms of trade, public spending, investment, commercial policy, have a significant influence on the real exchange rate level in the long-run.

12.2. Determinants of the real equilibrium exchange rate

Following Edwards (1993), we estimate the equilibrium real exchange rate level using a theoretical model where the simultaneous equilibrium of the current balance and the tradable good market is realized (see Emre Alper and Saglam, 2000).

Consider a small, open economy model with three goods – exportable (X), importable (M) and non-tradable (N). The economy involves consumers. The country produces non-tradable and exportable goods and consumes non-tradable and importable goods.

The country has a floating exchange rate system, with E denoting the nominal exchange rate in all transactions. This assumption may be surprising at first sight especially since numerous countries of our sample seem to have a fixed exchange rate. However we consider here a long-run horizon and estimate in the econometric part a long-run relationship. Of course the exchange rate can be fixed in short and mid terms but in the long-run countries must have a sufficient amount of currencies at their disposal to maintain the exchange rate which is not the case for most developing countries of our sample. As the nominal exchange rate will finally adjust here we directly suppose that it is flexible.

Let P_X and P_N be the prices of importable and non-tradable goods respectively. The world price of exportable goods is normalized to unity ($P_X^* = 1$), so the domestic price of exportable goods is $P_X = EP_X^* = E$. The world price of importable goods is denoted by P_M^*.

We define e_M and e_X as the domestic relative prices of importable and exportable goods with respect to non-tradable ones, respectively:

$$e_M = \frac{P_M}{P_N} \tag{12.1}$$

and

$$e_X = \frac{E}{P_N}. \tag{12.2}$$

Then the relative price of importable goods with respect to non-tradable ones is:

$$e_M^* = \frac{EP_M^*}{P_N}. \tag{12.3}$$

The country imposes tariffs on imports so that

$$P_M = EP_M^* + \tau, \tag{12.4}$$

where τ is the tariff rate.

The total output, Q, in the country is

$$Q = Q_X(e_X) + Q_N(e_X), \tag{12.5}$$

where $Q'_X > 0$ and $Q'_N < 0$.

Private consumption, C, is given by

$$C = C_M(e_M) + C_N(e_M), \tag{12.6}$$

where C_M and C_N are consumption on importable and non-tradable goods respectively, and $C'_M < 0$, $C'_N > 0$.

We define the real exchange rate as the relative price of tradable goods to non-tradable ones and denote it by e:

$$e = \alpha e_M + (1-\alpha)e_X = \frac{E(\alpha P^*_M + (1-\alpha)) + \alpha\tau}{P_N} \tag{12.7}$$

with $\alpha \in (0, 1)$.

Capital is perfectly mobile. The net foreign assets of the country are denoted by A. The country invests its net foreign assets at the international real interest rate r^*. The current account of the country in a given year is the sum of the net interest earnings on the net foreign assets and the trade surplus in foreign currency as the difference between the output of exportable goods and the total consumption of importable ones:

$$CA = r^*A + Q_X(e_X) - P^*_M C_M(e_M). \tag{12.8}$$

Change in the foreign currency reserves, R, of the country is then given by

$$.R = CA + KI, \tag{12.9}$$

where KI is the net capital inflows.

In the short and medium run, there can be departures from $.R = 0$, so that the country may gain or lose reserves. Current account is sustainable if the current account deficit plus the net capital inflows in the long run sum up to zero so that the official reserves of the country do not change. We then say that the economy is in external equilibrium if the sum of the current account balance and the capital account balance equal to zero, i.e.

$$r^*A + Q_X(e_X) - P^*_M C_M(e_M) + KI = 0, \tag{12.10}$$

$$C_N(e_M) + G_N = Q_N(e_X), \tag{12.11}$$

where G_N denotes public spending in non-tradable goods.

A real exchange rate is then said to be in equilibrium if it leads to external and internal equilibria simultaneously. From (12.9) and (12.10) it is possible to express the equilibrium exchange rate, e^*, as a function of P^*_M,

τ, r^*, A, KI and G_N, i.e.

$$e^* = e^*(P_M^*, \tau, r^*, A, KI, G_N). \tag{12.12}$$

The real exchange rate equilibrium level is thus a function of the terms of trade, commercial policy, the foreign interest rate, foreign capital flows, and public spending. The variables of Equation (12.12) are the fundamental of the real exchange rate in the long-run. An increase of public spending in non-tradable goods entails a real exchange rate appreciation, i.e. a deterioration of the country competitive position. A trade liberalization leads to a real depreciation of the domestic currency, i.e. an improvement of the country competitive position. An improvement of the trade balance entails a real exchange rate appreciation in the long-run. The effect of the terms of trade is ambiguous. On the one side, the terms of trade increase leads to a national income rise and hence to an expenditure rise and a real exchange rate appreciation. On the other, this increase generates a substitution effect and a real exchange rate depreciation. Elbadawi and Soto (1995) studied 7 developing countries and found that for three of them a terms of trade improvement entails of a real exchange rate appreciation, while for the four others, it led to a depreciation. Feyzioglu (1997) found that a terms of trade improvement entailed a real exchange rate appreciation in Finland.

12.3. Empirical investigation of the long term real exchange rate determinants

12.3.1. The econometric relationship to be tested and the data set

The theoretical model developed in Section 12.2 defines a long-run relationship between the real exchange rate and macroeconomic variables. The aim of this section is to test this relationship on panel data by taking explicitly the non-stationarity properties of the variables into account, and to identify the long term real exchange rate determinants. Indeed, before the development of econometric techniques adapted to non-stationary dynamic panels, previous studies on panel data implicitly supposed that the variables were stationary. This constitutes a serious limitation to their results given the considerable bias existing in this case on the parameter estimates when the non-stationarity properties of data are not taken into account. With the recent developments of econometrics it is henceforth possible to test stationarity on panel data as well as the degree of integration of the set of variables.[1]

[1] These tests are sufficiently well-known to exempt us from their formal presentation and for detailed discussions the reader will find references at the end of the paper.

Given the theoretical framework of Section 12.2, the cointegrating relationship to be tested between the real exchange rate and its fundamentals can be written as:

$$\begin{aligned} tcr_{it} &= \alpha_{1i} + \beta_{1i} te_{it} + \beta_{2i} ouv_{it} + \beta_{3i} fdi_{it} + \beta_{4i} g_{it} + \beta_{5i} inv_{it} \\ &\quad + \beta_{6i} gdp_{it} + \varepsilon_{it}, \quad i = 1, 2, \ldots, N \text{ and } t = 1, 2, \ldots, T \end{aligned} \tag{12.13}$$

with:

- *e*: the logarithm of the real exchange rate quoted to incertain,
- *te*: the logarithm of the terms of trade,
- *ouv*: the logarithm of trade policy,
- *fdi*: the logarithm of foreign direct investments flows (FDI, in percentage of GDP),
- *g*: the logarithm of the share of public spending in the GDP,
- *inv*: the logarithm of domestic investment (in percentage of GDP),
- *gdp*: the logarithm of GDP per capita.

We consider a sample of 45 developing countries, divided into three groups according to geographical criteria: *Africa* (21 countries: Algeria, Benin, Burkina Faso, Burundi, Cameroon, Congo, the democratic Republic of Congo, Ivory Coast, Egypt, Ethiopia, Gabon, Gambia, Ghana, Guinea Bissau, Kenya, Mali, Morocco, Mozambique, Niger, Senegal, Tunisia), *Latin America* (17 countries: Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Peru, Paraguay, Uruguay, Venezuela) and *Asia* (7 countries: Bangladesh, Indonesia, South Korea, India, Malaysia, the Philippines, Thailand).

The sample period is based on data availability and it covers 16 years for Africa (from 1980 to 1996), 23 years for Latin America (from 1973 to 1996) and 21 years for Asia (from 1975 to 1996). All the data are annual and are extracted from the World Bank data base for the fundamental[2] and from the French database of the CEPII (CHELEM) for the real exchange rate.[3] The real exchange rate is calculated as the ratio of the consumer

[2] As pointed out by a referee, although we consider reliable official data sets, they do not accurately represent developing economies where black markets and corruption are major actors. Solving this issue seems however impossible as of now and we therefore acknowledge this limitation in our conclusions.

[3] The CHELEM database has been used here for the TCR because our work is an extension of another study published in 2004 in the n° 97-1 issue of *Economie Internationale* on the investigation of the validity of purchasing power parity (PPP) theory for developing countries. In that study we used data extracted from on the TCR extracted from the CHELEM French database and showed that the two versions of PPP (weak and strong)

price index in the United States (CPI) to that of the considered country multiplied by the nominal exchange rate with regard to the US Dollar and an increase implies a depreciation. The terms of trade are calculated as the ratio of export price index to import price index of the considered country. Domestic investment is calculated as the ratio of gross investment at constant prices to the sum of private consumption, government consumption, and gross investment, all at constant prices.

Let us underline that the unavailability of data for some macroeconomic variables led us to proceed to some approximations. The first one is related to public spending in non-tradable goods: as we cannot decompose them into tradable and non-tradable goods, we used the global public spending share (in value) in GDP (in value) as a proxy. The second one concerns trade policy. Generally, in literature, the openness degree of the economy is approximated by the share of foreign trade in GDP (in value). This approximation justifies itself by the fact that *ceteris paribus*, a greater tradable liberalization allows to intensify trade and the convergence of prices. In our case we used the share of total imports (in value) in total domestic spending (in value).

Long-run capital movements are approximated by foreign direct net flows (FDI). This choice justifies itself by the fact that contrary to other financial flows, the FDI are related to output motivations and are therefore more stable.

Per capita income is used as a proxy to measure the Balassa–Samuelson effect (cf. Balassa, 1964). We expect the coefficient of per capita income to be negative since economic development comes along with an increasing gap between the relative productivity in the tradable sector, which leads to a real exchange rate appreciation.

Note that the cointegration coefficients are estimated by the fully modified least square method (Fmols), developed by Pedroni (2000). The advantage of this method with regard to the standard MCO is that it corrects distortions related to the correlation between regressors and residuals and that it is less sensitive to possible bias in small size samples (cf. Pedroni, 2000).

12.3.2. Econometric results and their economic interpretation

The analysis first step is simply to look at the data univariate properties and to determine their integratedness degree. In this section, we implement

were not relevant to describe the long-run behavior of the TCR in Africa, Asia and South America. Our goal here is to refine this analysis and we examine therefore explicitly the long-run determinants of the TCR for these three groups of countries using data from the same source.

Table 12.1. Equilibrium real exchange rate estimation

	te	*ouv*	*fdi*	*g*	*inv*	*gdp*	ADF-stat.	*p* val.	Bootstrap distribution		
									1%	5%	10%
									−3.02	−2.19	−1.72
Africa											
Coeff.	−0.56	0.16	−0.06	0.05	0.17	−0.07					
t-stat.	−8.58	2.38	−2.76	2.92	3.04	−3.62	−5.91	0.00			
Latin America											
Coeff.	*ns*	0.09	−0.02	−0.10	0.17	−0.23					
t-stat.	*ns*	2.97	−3.21	−2.43	3.04	−3.35	−3.82	0.00			
Asia											
Coeff.	−0.53	0.39	−0.07	0.13	0.37	−0.39					
t-stat.	−2.94	11.01	−4.58	3.53	2.11	−10.08	−12.16	0.00			

Note. The bootstrap is based on 2000 replications.

three panel data unit root tests (Im *et al*., 1997, 2003; Choi, 2002; Moon and Perron, 2003) in order to investigate the robustness of our results.

First, we used the test proposed by Im *et al*. (1997, 2003, hereafter IPS) that has been widely implemented in the empirical research due to its rather simple methodology and alternative hypothesis of heterogeneity. This test assumes cross-sectional independence among panel units, but allow for heterogeneity of the form of individual deterministic effects (constant and/or linear time trend) and heterogeneous serial correlation structure of the error terms. Table A12.1 in the appendix reports the results of the IPS's test and indicates that the null hypothesis of unit-root cannot be rejected at the 5% level for all series.

However, as shown by several authors (including O'Connell, 1998; Banerjee *et al*., 2004a, 2004b), the assumption of cross-sectional independence on which the asymptotic results of the IPS's procedure relies (as actually most panel data unit root tests of "the first generation" including Maddala and Wu, 1999; Levin and Lin, 1993; Levin *et al*., 2002) is often unrealistic and can be at odds with economic theory and empirical results. Besides, as shown in two simulation studies by Banerjee *et al*. (2004a, 2004b) if panel members are cross-correlated or even cross-sectionally cointegrated, all these tests experience strong size distortions and limited power. This is analytically confirmed by Lyhagen (2000) and Pedroni and Urbain (2001).

For this reason, panel unit root tests relaxing the assumption of cross-sectional independence have recently been proposed in the literature including Choi's (2002), Bai and Ng's (2003), Moon and Perron's (2003), Pesaran's (2003) and Phillips and Sul's (2003) tests. We have decided to investigate the presence of a unit-root using two tests of "the second generation", the test proposed by Choi (2002), and that by Moon and Perron's (2003), to whom we refer the reader for further details. This last test in particular, seems to show "good size and power for different values of T and N and model specifications", according to the Monte Carlo experiments by Gutierrez (2003). The results reported in Tables A12.2 and A12.3 in the appendix indicate that the null hypothesis of unit-root cannot be rejected by the two tests at the 5% level for our seven series, for Africa, Latin America and Asia, hence supporting the first results given by the IPS's test. Furthermore, tests on the series in first differences confirm the hypothesis of stationarity. We therefore conclude that the real exchange rate and its potential determinants expressed in level are all integrated of order 1, independently of the panel unit-root tests considered, which tend to prove that the non-stationarity property of our macro-economic series is a robust result.

Afterwards, having confirmed the non-stationarity of our series, it is natural to test the existence of a long-run relationship between the real exchange rate and its determinants. Table 12.1 reports the results of the panel data cointegration tests developed by Pedroni (1999, 2004) both using conventional (asymptotic) critical values given in Pedroni (1999) and bootstrap critical values.[4] Indeed, the computation of the Pedroni statistics assumes cross-section independence across individual i, an assumption that is likely to be violated in many macroeconomic time series (see Banerjee *et al.*, 2004a, 2004b), including in our study. In order to take into account the possible cross-section dependence when carrying out the cointegration analysis, we have decided to compute the bootstrap distribution of Pedroni's test statistics and have generated in this way data specific critical values. Note that as in Banerjee and Carrion-i-Silvestre (2004), we have of course not used the seven statistics proposed by Pedroni (1999, 2004) (to test the null hypothesis of no cointegration using single equation methods based on the estimation of static regressions). These statistics can also be grouped in either parametric or non-parametric statistics, depending on the way that autocorrelation and endogeneity bias is accounted for. In our study, we are only concerned with the parametric version of the statistics, i.e. the normalized bias and the pseudo t-ratio statistics and more precisely with the ADF test statistics. These test statistics

[4] Let us underline that as we implement a one sided test a calculated statistic smaller than the critical value leads to the rejection of the null hypothesis of absence of a cointegration relationship between the variables. Note also that β_j represents the average of the estimated β_{ij} for j varying from 1 to 6 (cf. Equation (12.13)).

are defined by pooling the individual tests, so that they belong to the class of between dimension test statistics (cf. Pedroni, 1999, 2004 for further details).

It is also important to notice that, as stressed by Banerjee and Carrion-i-Silvestre (2004), some cautions about the method that is used to bootstrap cointegration relationships are required, since not all available procedures lead to consistent estimates. In this regard, we have followed Phillips (2001), Park (2002), and Chang *et al*. (2002), and we have decided to use sieve bootstrap using the modified version of the sieve bootstrap[5] described in Banerjee *et al*. (2004a, 2004b).[6]

Using both the conventional (asymptotic) critical values (−1.65 at 5%) calculated under the assumption of cross-section independence (reported in Pedroni, 1999, and extracted from the standard Normal distribution), and our bootstrap critical value (−2.19 at 5%, valid if there is some dependence amongst individuals), the null hypothesis of no cointegration is always rejected by test statistics. Therefore, we conclude that a long-run relationship exists between the real exchange rate and its fundamentals for our three sets of countries (in Africa, Latin America and Asia).

Empirical results (cf. β_1) confirm that an improvement of the terms of trade entails a real exchange rate appreciation in Africa and in Asia, which means that the wealth effect dominates the substitution effect. Furthermore, the elasticity of the real exchange rate with respect to the terms of trade is compatible with previous studies. The difference between the economic structures of the two groups of countries partially explains the difference of response of real exchange rates to a shock on the terms of trade (an improvement of 10% of the terms of trade entails an appreciation of 5.6% in Africa and 5.3% in Asia). The absence of the effect of the terms of trade on the real exchange rate in Latin America confirms that the wealth effect compensates for the substitution effect.

Negative coefficients (β_2) for the three groups of countries suggest that trade liberalization is accompanied with a real exchange rate depreciation. The elasticity is different for the three groups of countries: it is of 0.16 in Africa, 0.39 in Asia and 0.09 in Latin America. Nevertheless, this elasticity remains relatively low for these countries in comparison to the previous results of literature (Elbadawi and Soto, 1995; Baffes *et al.*, 1999). A possible explanation is that the estimated coefficients are averages of individual coefficients.

[5] We are very grateful to Banerjee and Carrion-i-Silvestre for providing us their Gauss codes.

[6] For a detailed discussion the reader will find references at the end of the paper.

For these three groups the cointegration coefficients of the FDI confirm the theoretical predictions. The estimated coefficient (β_3) is negative, implying that a capital flow increase entails a domestic spending rise and a reallocation of output factors towards the non-tradable goods sector; the long-run demand increase of non-tradable goods entails a real exchange rate appreciation. Furthermore, the coefficients are very close for the three groups of countries. Indeed, an increase of 1% of foreign investments flows leads to an average real exchange rate appreciation of 0.05%.

The effect of public spending on real exchange rates (β_4) is different for the three groups of countries. Indeed, estimations indicate that an increase of public spending entails a real exchange rate appreciation in Latin America and a depreciation in Asia and Africa. According to theoretical predictions the coefficient must be negative given that the increase of the global demand of non-tradable goods entails an increase of their price. The positive coefficient in Asia and Africa can reflect a strong eviction effect which induces a fall in private non-tradable goods demand. If public spending is extensive in tradable goods, an expansionist budget policy entails a tax increase or/and an interest rate rise, which reduces the private demand of non-tradable goods. The fall in demand then entails a price decrease and hence a real exchange rate depreciation (cf. Edwards, 1989). The effect of public spending on the real exchange rate in Latin America and in Asia is comparable and relatively higher than in Africa.

An increase of 10% on the share of domestic investments entails an average depreciation of 1.7% in Africa and in Latin America and of 3.7% in Asia (coefficient β_5). This result is compatible with that of Edwards (1989) which also found a low elasticity (of 7%) for a group of 12 developing countries. Indeed, an increase of investments often leads to an increase of non-tradable goods spending and hence to a decrease of the relative price of non-tradable goods.

The per capita GDP also contributes to the long-run variations of the real exchange rate for the three groups of countries. The coefficient (β_6) is negative, which implies that economic development is accompanied by a real exchange rate appreciation (Balassa–Samuelson effect). The effect of economic development on the long-run evolution of the real exchange rate is relatively low in Africa. Indeed, an increase of 1% of per capita GDP entails a real exchange rate appreciation of only 0.07%. On the other hand, this effect is relatively high in Asia and Latin America since real exchange rate appreciates respectively of 0.39% and 0.23% for these countries following an increase of 1% of the per capita GDP.

Finally, notice that in Africa and in Asia external factors (openness degree and terms of trade) contribute most to the long-run dynamics of the real exchange rate; internal demand also plays an important role in Asia. In Latin America on the other hand, external factors seem to have a relatively limited effect on the equilibrium real exchange rate, the economic development (GDP per capita) having on the contrary an important role.

12.4. Conclusion

The aim of this paper was to identify the determinants of the equilibrium real exchange rate for developing countries. On the basis of theoretical approaches used in literature, we have exposed a simple theoretical model which describes the interaction between some macroeconomic variables and the equilibrium real exchange rate level. Then, this model has been estimated by recent non-stationary panel data techniques. We have in particular used the panel data integration tests proposed by Im *et al.* (1997, 2003), Choi (2002) and Moon and Perron's (2003) (the last two tests relaxing the assumption of cross-section independence across individual i which is rather unrealistic in applied research), as well as the panel data co-integration framework developed by Pedroni (1999, 2004) and generalized by Banerjee and Carrion-i-Silvestre (2004). In particular, following Banerjee *et al.* (2004a, 2004b), we have bootstrapped the critical values of Pedroni's cointegration tests under the assumption of cross-section dependence. These recent advances in the econometrics of non-stationary panel methods have enabled us to put in evidence the existence of several sources of impulsions influencing the real exchange rate in the long-term in Africa, Latin America and Asia.

Our investigations show that an improvement of the terms of trade, an increase of per capita GDP and of capital flows entail a long-run appreciation of the real exchange rate. On the other hand, an increase of domestic investment and of the openness degree of the economy entails a real exchange rate depreciation; the effect of public spending increase being ambiguous.

Our results confirm that the real exchange rate depends on the economic specificities of each country. In other words, we don't have a fixed and general norm but, for each economy, the real exchange rate trajectory depends on its development level, on the way economic policy is conducted, and on its position on the international market. Besides, the variations of the real exchange rate do not necessarily reflect a disequilibrium. Indeed, equilibrium adjustments related to fundamental variations can also generate real exchange rate movements.

Notice finally that the non-stationary panel data econometric approach applied here to 45 countries does not directly allow us to determine the over (under) evaluations for each country individually.

Acknowledgements

We would like to thank Anindya Banerjee and Josep Lluís Carrion-i-Silvestre for providing us some Gauss codes for the modified version of the sieve bootstrap described in Banerjee *et al.* (2004a, 2004b) for Pedroni's test statistics as well as in Choi for the Gauss Code of his panel data unit-root test available at http://ihome.ust.hk/inchoi/. We are also very grateful to Badi H. Baltagi and to an anonymous referee for helpful remarks and suggestions on an earlier draft of this paper. The usual disclaimer applies.

Appendix A12. Panel unit-root test results for developing countries[7]

*Table A12.1. Results of Im et al. (1997, 2003) test**

	Level		First difference	
	Constant	Constant and trend	Constant	Constant and trend
Real exchange rate				
Africa	−1.34	−1.6	−2.30	−2.38
Latin America	−0.23	−1.43	−3.32	−4.32
Asia	−0.32	−1.65	−2.54	−2.12
GDP per capita				
Africa	−0.09	−1.60	−2.30	−2.38
Latin America	−0.12	−1.43	−2.21	−2.54
Asia	−0.19	−1.45	2.31	−3.45
Terms of trade				
Africa	−0.66	−0.17	−7.77	−5.72
Latin America	−0.32	−0.43	−5.45	−5.21
Asia	−0.36	−0.32	−5.47	−6.32
Openness degree				
Africa	−0.55	−0.63	−2.33	−7.77
Latin America	−0.43	−0.98	−3.23	−6.47
Asia	−0.12	−0.43	−2.54	−3.34
Public spending				
Africa	−0.79	−1.79	−3.45	−4.05
Latin America	−1.32	−1.12	−2.31	−3.21
Asia	−0.86	−1.68	−3.32	−4.65
Foreign direct investments				
Africa	−0.19	−1.62	−2.63	−4.35
Latin America	−0.12	−1.43	−2.12	−5.22
Asia	−0.21	−1.42	−2.55	−3.21
Domestic investments				
Africa	−0.23	−1.14	−3.89	−3.23
Latin America	−0.41	−1.21	−3.32	−4.23
Asia	−1.32	−1.35	−3.21	−4.67

*As this is one-sided tests, the critical value is −1.65 (at the 5% level) and for unit-root to exist the calculated statistics must be larger than −1.65.

[7] All variables are expressed in logarithms.

Table A12.2. Results of Choi's (2002) test*

	P_m statistic	Z statistic	L^* statistic
Real exchange rate			
Africa	0.177**	0.356	0.194
Latin America	0.103	0.172	0.224
Asia	0.07	0.08	0.06
GDP per capita			
Africa	0.091	0.321	0.159
Latin America	0.061	0.311	0.05
Asia	0.798	0.987	0.975
Terms of trade			
Africa	0.128	0.071	0.062
Latin America	0.054	0.081	0.056
Asia	0.321	0.421	0.452
Openness degree			
Africa	0.254	0.321	0.341
Latin America	0.051	0.074	0.047
Asia	0.562	0.547	0.412
Foreign direct investments			
Africa	0.112	0.125	0.185
Latin America	0.045	0.568	0.098
Asia	0.256	0.341	0.387
Domestic investments			
Africa	0.098	0.093	0.150
Latin America	0.045	0.105	0.07
Asia	0.121	0.231	0.192

*Note that the P_m test is a modification of Fisher's (1932) inverse chi-square tests and rejects the null hypothesis of unit-root for positive large value of the statistics, and that the L^* is a logit test. The tests (Z and L^*) reject the null for large negative values of the statistics. The P, Z and L^* tests converge under the null to a standard normal distribution as $(N, T \to \infty)$, cf. Choi's (2002) for further details.

**All figures reported in Table A12.2 are P-values.

Table A12.3. Results of Moon and Perron's (2003)[*]*

	$t * a$	$t * b$
Real exchange rate		
Africa	0.153[**]	0.124
Latin America	0.421	0.342
Asia	0.182	0.147
GDP per capita		
Africa	0.921	0.752
Latin America	0.354	0.247
Asia	0.165	0.198
Terms of trade		
Africa	0.051	0.061
Latin America	0.042	0.067
Asia	0.321	0.258
Openness degree		
Africa	0.147	0.189
Latin America	0.159	0.325
Asia	0.487	0.362
Foreign direct investments		
Africa	0.321	0.273
Latin America	0.092	0.121
Asia	0.043	0.051
Domestic investments		
Africa	0.484	0.517
Latin America	0.397	0.377
Asia	0.071	0.0521

[*] The null hypothesis of the two tests proposed by Moon and Perron (2003) is the unit-root for all panel units. Under the null H_0, MP show that for $(N, T \to \infty)$ with $N/T \to 0$ the statistics $t * a$ and $t * b$ have a standard normal distribution.
[**] All figures reported in Table A12.3 are P-values.

References

Baffes, J., Elbadawi, I., O'Connel, S. (1999), "Single-equation estimation of the equilibrium real exchange rate", in: Hinkle, L., Montiel, P., editors, *Exchange Rate and Measurement for Developing Countries*, Oxford University Press, pp. 405–464.

Bai, J., Ng, S. (2003), "A PANIC attack on unit roots and cointegration", *Econometrica*, submitted for publication.

Balassa, B. (1964), "The purchasing power parity doctrine: a reappraisal", *Journal of Political Economy*, Vol. 72 (6), pp. 584–596.

Banerjee, A. (1999), "Panel data units and cointegration: an overview", *Oxford Bulletin of Economics and Statistics*, Vol. 61 (3), pp. 607–629.

Banerjee, A., Carrion-i-Silvestre, J.L. (2004), "Breaking panel data cointegration", Preliminary draft, October, downloadable at http://www.cass.city.ac.uk/conferences/cfl/CFLPapersFinal/Banerjee%20and%20Carrion-i-Silvestre%202004.pdf.

Banerjee, A., Marcellino, M., Osbat, C. (2004a), "Testing for PPP: should we use panel methods?", *Empirical Economics*, submitted for publication.

Banerjee, A., Marcellino, M., Osbat, C. (2004b), "Some cautions on the use of panel methods for integrated series of macro-economic data", *Econometrics Journal*, submitted for publication.

Chang, Y., Park, J.Y., Song, K. (2002), "Bootstrapping cointegrating regressions", Mimeo, Rice University.

Choi, I. (2002), "Combination unit root tests for cross-sectionally correlated panels", Mimeo, Hong Kong University of Science and Technology.

Cottani, J.A., Cavallo, F., Khan, M.S. (1990), "Real exchange rate behavior and economic performance in LDCs", *Economic Development and Cultural Change*, Vol. 39 (3), pp. 61–76.

Drine, I., Rault, C. (2002), "Do panel data permit to rescue the Balassa–Samuelson hypothesis for Latin American countries?", *Applied Economics*, Vol. 35 (3), pp. 351–361.

Drine, I., Égert, B., Lommatzsch, K., Rault, C. (2003), "The Balassa–Samuelson effect in Central and Eastern Europe: myth or reality?", *Journal of Comparative Economics*, Vol. 31 (3).

Edwards, S. (1989), *Real Exchange Rates, Devaluation and Adjustment: Exchange Rate Policy in Developing Countries*, MIT Press, Cambridge, MA.

Edwards, S. (1993), "Openness, trade liberalization, and growth in developing countries", *Journal of Economic Literature, American Economic Association*, Vol. 31 (3), pp. 1358–1393.

Elbadawi, I., Soto, R. (1995), "Capital flows and equilibrium real exchange rate in Chile", Policy Research Working Paper No. 1306, Banque mondiale, Washington, DC.

Emre Alper, C., Saglam, I. (2000), "Equilibrium level of the real exchange rate and the duration and magnitude of the misalignments for Turkey", *Proceedings of the Middle East Economic Association*, in conjunction with Allied Social Sciences Association in Boston, MA, USA, January 7–9.

Feyzioglu, T. (1997), "Estimating the equilibrium real exchange rate: an application to Finland", IMF Working Paper 97/109, IMF, Washington, DC, August.

Fisher, R.A. (1932), *Statistical Methods for Research Workers*, Oliver and Boyd, London.

Gutierrez, L. (2003), "Panel unit roots tests for cross-sectionally correlated panels: a Monte Carlo comparison", Econometrics 0310004, Economics Working Paper Archive at WUSTL.

Im, K.S., Pesaran, M.H., Shin, Y. (1997), "Testing for unit roots in heterogeneous panels", Discussion Paper, University of Cambridge, June.

Im, K.S., Pesaran, M.H., Shin, Y. (2003), "Testing for unit roots in heterogeneous panels", *Journal of Econometrics*, Vol. 115, pp. 53–74.

Levin, A., Lin, C.F. (1993), "Unit root tests in panel data, asymptotic and finite sample properties", U.C. San Diego Working Paper.

Levin, A., Lin, C.F., Chu, C.J. (2002), "Unit root tests in panel data: asymptotic and finite-sample properties", *Journal of Econometrics*, Vol. 108, pp. 1–24.

Lyhagen, J. (2000), "Why not use standard panel unit root test for testing PPP", Mimeo, Stockholm School of Economics.

Maddala, G., Wu, S. (1999), "A comparative study of unit root tests and a new simple test", *Oxford Bulletin of Economics and Statistics*, Vol. 61, pp. 631–652.

Mac Donald, R., Ricci, L. (2003), "Estimation of the equilibrium real exchange rate for South Africa", IMF Working Paper, 2003/44, IMF, Washington, DC.

Moon, H.R., Perron, B. (2003). "Testing for a unit root in panels with dynamic factors", Mimeo, University of Southern California Law School.

O'Connell, P.G.J. (1998), "The overvaluation of purchasing power parity", *Journal of International Economics*, Vol. 44, pp. 1–19.

Park, J.Y. (2002), "An invariance principle for sieve bootstrap in time series", *Econometric Theory*, Vol. 18, pp. 469–490.

Pedroni, P. (1999), "Critical values for cointegrating tests in heterogeneous panels with multiple regressors", *Oxford Bulletin of Economics and Statistics*, Vol. 61 (Supplement), pp. 653–670.

Pedroni, P. (2000), "Fully modified OLS for heterogeneous cointegrated panels", *Advances in Econometrics*, Vol. 15, pp. 93–130.

Pedroni, P. (2004), "Panel cointegration; asymptotic and finite sample properties of pooled time series tests with an application to the purchasing power parity hypothesis", *Econometric Theory*, Vol. 20, pp. 597–625.

Pedroni, P., Urbain, J.-P. (2001), "Cross member cointegration in non-stationary panels", Mimeo, Universtiteit Maastricht.

Pesaran, M.H. (2003), "A simple panel unit root test in the presence of cross section dependence", Mimeo, Cambridge University.

Phillips, P.C.B. (2001), "Bootstrapping spurious regressions", Cowles Foundation Discussion Paper 1330, Yale.

Phillips, P.C.B., Sul, D. (2003), "Dynamic panel estimation and homogeneity testing under cross-section dependence", *Econometrics Journal*, Vol. 6, pp. 217–259.

Strauss, J. (1999), "Productivity differentials, the relative price of non-tradables and real exchange rates", *Journal of International Money and Finance*, Vol. 18 (3), pp. 383–409.

Xiaopu, Z. (2002), "Equilibrium and misalignment: an assessment of the RMB exchange rate from 1978 to 1999", Center for Research on Economic development and Policy Reform Working Paper, No. 127.

Further reading

Drine, I., Rault, C. (2004), "La théorie de la parité du pouvoir d'achat est-elle vérifiée pour les pays développés et en développement? Un ré-examen par l'économétrie des panels non-stationnaires", *Economie Internationale*, No 97-1er trimestre.

Edwards, S., Savastano, M.A. (1999), "Exchange rates in emerging economies: what do we know? What do we need to know?", NBER Working Papers 7228, National Bureau of Economic Research, Inc.

This page intentionally left blank

Panel Data Econometrics
B.H. Baltagi (Editor)
© 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S0573-8555(06)74013-5

CHAPTER 13

Employee Turnover: Less is Not Necessarily More?[*]

Mark N. Harris[a], Kam Ki Tang[b] and Yi-Ping Tseng[c,**]

[a]Department of Econometrics and Business Statistics, Monash University, Australia
E-mail address: mark.harris@buseco.monash.edu.au
[b]School of Economics, University of Queensland, Australia
E-mail address: kk.tang@uq.edu.au
[c]Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Australia
E-mail address: y.tseng@unimelb.edu.au

Abstract

Theoretical studies have suggested firm specific human capital and job matching as the major, but opposite, mechanisms through which employee turnover affects labour productivity. This study finds that the former dominates when turnover is high, while the latter dominates when turnover is low. The optimal turnover rate that maximises productivity is about 0.22 per annum. Bringing the observed turnover rates in the sample to the optimal level increases the average productivity by 1.1 per cent. The large gap between the observed and the optimal rate could be explained by the lack of decision coordination between agents in labour markets.

Keywords: employee turnover, productivity, firm specific human capital, job matching, panel data, coordination

JEL classifications: J41, J63

[*] This paper is the result of work being undertaken as part of a collaborative research program entitled: “The Performance of Australian Enterprises: Innovation, Productivity and Profitability”. The project is generously supported by the Australian Research Council and the following collaborative partners: the Australian Tax Office, the Commonwealth Office of Small Business, IBIS Business Information Pty Ltd, the Productivity Commission, and the Victorian Department of State Development. The views expressed in this chapter, however, represent those of the authors and are not necessarily shared by the collaborative partners.

[**] Corresponding author.

13.1. Introduction

It is widely acknowledged in the business community that human resources are an invaluable firm asset (see, for example, Business Asia, 1999; Business Times, 2000). Therefore, it is logical to assume that the flow of this valuable asset – employee turnover – will play a crucial role in firm performance. Indeed, firms (and employees) are burdened with turnover problems in both good and adverse economic climates. During economic upturns, employee churning represents one of the greatest difficulties in business management. For instance, during the "new economy" boom in the U.S., nearly a quarter of workers were reported to have average tenure of less than a year (Economist, 2000).[1] On the other hand, during economic downturns, trimming operating costs through job retrenchment in order to maintain a firm's share value is a typical phenomenon. Nevertheless, downsizing is not a painless option for firms, as they are likely to suffer adverse consequences, such as low levels of morality and loyalty amongst the remaining staff. Moreover, firms also bear the risk of not being able to quickly re-establish the workforce should the economy rebound more swiftly than anticipated.

As a consequence, employee turnover has been extensively researched across a number of disciplines, including: psychology; sociology; management; and economics. Each discipline has its own focus and, accordingly, employs different research methodologies. Psychologists and sociologists, for example, are generally interested in the motivations behind quitting, such as job satisfaction, organisational commitment and job involvement (Carsten and Spector, 1987; Muchinsky and Tuttle, 1979). Empirical work in these fields typically involves case studies using survey data of individual firms or organisations.

In the discipline of management study, high staff turnover has been of great and continuous concern (as typified by Mok and Luk, 1995, and the symposium in *Human Resource Management Review*, Vol. 9 (4), 1999). Similar to the practice in psychology and sociology, researchers heavily draw on event, or case, studies. While reducing employee turnover is a managerial objective for some firms, the converse is true for others. For example, legal restrictions and obligations in recruitment and dismissal could prohibit firms from maintaining a flexible workforce size, a situation more common in unionised sectors (Lucifora, 1998). The industrial reforms and privatisation in many developed nations were aimed, at least in part, at increasing the flexibility of labour markets.

[1] High-tech industries as well as the low-tech ones, such as retailing, food services and call centres, experienced the problem.

In contrast, economists focus mainly on the implications of turnover on unemployment. A strand of matching theories has been developed extensively to explain equilibrium unemployment, wages and vacancies (Lucas and Prescott, 1974; Lilien, 1982). National aggregate time series data are typically employed in this line of research. For recent surveys on matching theories and their applications see Petrongolo and Pissarides (2001) and the symposium in *Review of Economic Studies*, Vol. 61 (3), 1994.

Despite turnover being considered crucial to human resource management and production, there is little quantitative research on the effect of turnover on labour productivity (hereafter "productivity" unless specified otherwise).[2] This omission is possibly due to the lack of firm level data on *both* production and turnover. Moreover, firm level data are typically restricted to individual organisations, prohibiting researchers from drawing general conclusions.[3] Utilising a recently released firm-level panel data set, based on the Australian Business Longitudinal Survey (BLS), this paper is therefore able to provide a new dimension to the literature. The BLS data provide an objective measure of value-added, which is comparable across firms operating in a broad spectrum of industries. Conditional on firm level factor inputs and other firm characteristics, the impacts of employee turnover on productivity are investigated. The results suggest that employee turnover has a statistically significant and quantitatively large, but more importantly, non-linear effect on productivity. From the results it is possible to estimate the optimal turnover rate – the rate that maximises productivity, keeping other factors constant – which was found to be around 0.22 per annum. As the employee turnover rate is defined here as the average of total number of employees newly recruited and departed within a period, relative to the average number of employees over the period, the highest productivity is where about 22 per cent of total employees changed over the one-year period. The estimated optimal rate is much higher than that typically observed in the sample (the median turnover rate is about 14 per cent). Using a theoretical model, it is shown that the lack of coordination between agents in labour markets can lead them choosing

[2] McLaughlin (1990) examines the relationship between turnover type (quit or layoff) and economy-wide general productivity growth, but not productivity of individual firms. Shepard *et al.* (1996) make use of survey data to estimate the total factor productivity of the pharmaceutical industry; nevertheless, their study is only concerned with the effect of flexible working hours and not turnover.

[3] For instance, Borland (1997) studies the turnover of a medium-size city-based law firm, Iverson (1999) examines voluntary turnover of an Australian public hospital, and Glenn *et al.* (2001) focus on major league baseball in the U.S. However, all three studies do not cover the production aspect of the examined organisation.

a turnover rate far below the optimal level. The intuition is that the possibility for an employer to find more productive staff (or for an employee to find a more rewarding job) is related to the rate of job-worker separations in other firms. Without sufficient information about the intended decisions of others, agents will make changes at sub-optimal rates.

The empirical results also suggest that if firms bring their turnover rates to the optimal level, average productivity will increase by just over 1 per cent. These results have clear policy implications. For instance, if the observed turnover rate is substantially below the estimated optimal rate and *if* institutional rigidity in the labour market is the main cause of that, deregulation may be warranted.

The rest of the paper is structured as follows. Section 13.2 reviews two main contending theories about the linkage between employee turnover and productivity, and formulates the concept of the optimal turnover rate. In Section 13.3 the econometric model and the data are briefly described. Section 13.4 presents the empirical results and Section 13.5 concludes. Appendix A13 provides details of the data, including summary statistics. Appendix B13 presents a theoretical model to account for the empirical findings.

13.2. Theories of employee turnover and productivity

There are two main theories on how employee turnover can affect productivity. Firstly, there is the firm specific human capital (FSHC) theory, pioneered by Becker (1975). This asserts that if firms need to bear the cost of training, their incentives to provide staff training will be lowered by high turnover rates. The incentive will be even weaker when firm specific and general training are less separable, as employees have lower opportunity costs of quitting (Lynch, 1993). Consequently, productivity falls as turnover increases. Even if FSHC is bred through learning-by-doing, its accumulation remains positively related to employees' tenure. As a result, a higher turnover rate will still lead to lower productivity.

In addition to the direct loss of human capital embodied in the leavers, there are other negative impacts of turnover on productivity. Besides the output forgone during the vacant and training period, the administrative resources used in separation, recruitment and training could have been invested in other aspects of the production process.[4] Moreover, high employee turnover could adversely affect the morale of the organisation.

[4] It has been reported that the cost of losing an employee is between half to one and a half times the employee's annual salary (Economist, 2000).

Using a controlled experiment, Sheehan (1993) records that the leavers alter the perceptions of the stayers about the organisation and therefore negatively affect its productivity. As a consequence, warranted (from an employer's perspective) but involuntary job separation could trigger unwarranted voluntary employee departure – a snowball effect.[5]

On the opposite side of the debate, is the job matching theory established by Burdett (1978) and Jovanovic (1979a, 1979b). The key insight of this theory is that firms will search for employees and job seekers will search for firms until there is a good match for both parties. However, the conditions for an optimal matching may change over time, leading to continuous reallocation of labour. For instance, a firm that has upgraded its production technology will substitute skilled for unskilled labour (Ahn, 2001). Moreover, established firms also need 'new blood' to provide fresh stimulus to the *status quo*. On the other hand, a worker who has acquired higher qualifications via education, training, or learning-by-doing may seek a better career opportunity.

Regular employee turnover helps both employers and employees avoid being locked in sub-optimal matches permanently. For instance, the estimated cost of a poor hiring decision is 30 per cent of the first year's potential earning and even higher if the mistake is not corrected within six months, according to a study by the U.S. Department of Labor (cited in Abbasi and Hollman, 2000).

Another factor that compounds the effect of turnover on productivity is knowledge spillover between firms (Cooper, 2001). Knowledge spillover is more significant if human capital is portable across firms or even industries. Megna and Klock (1993) find that increasing research input by one semi-conductor firm will increase the productivity of rival firms due to labour migration. Finally, Borland (1997) suggests that involuntary turnover can be used as a mechanism to maintain employees' incentives. In short, matching theory suggests that higher turnover aids productivity.

Although FSHC theory and job matching theory suggest opposite effects of turnover on productivity, one does not necessarily invalidate the other. In fact, there is empirical evidence supporting the coexistence of both effects, albeit the effect of FSHC appears to dominate (Glenn *et al.*,

[5] During the economic downturn in the U.S. in 2001, executives in Charles Schwab and Cisco were reportedly cutting down their own salaries and setting up charitable funds for laid off staff in order to maintain the morale of the remaining employees (Economist, 2001). Both companies' efforts were apparently well received. Fortune (2002) ranked Cisco and Charles Schwab as the 15th and 46th best companies to work for in 2001, respectively, despite Cisco was reported laying off 5,500 staff while Charles Schwab 3,800 staff.

2001). The two theories essentially answer the question of how to balance the stability and flexibility of the labour force. It is the contention here, that given FSHC and job matching have opposite effects on productivity, there is a distinct possibility that a certain turnover rate will maximise productivity. A scenario, in which such an optimal turnover rate exists, is where productivity is a non-linear – specifically quadratic concave function, of turnover.

13.3. Data, empirical model and estimation method

13.3.1. Business longitudinal survey

The BLS is a random sample of business units selected from the Australian Bureau of Statistics business register for inclusion in the first year of the survey. The sample was stratified by industry and firm size. The sample was selected with the aim of being representative of all businesses (excluding government agents, public utilities and public services). The focus is on a balanced panel of small and medium sized businesses. After excluding businesses with deficient data records, 2,435 businesses are left in our sample. Summary statistics and variable definitions are presented in Appendix A13.

This data source is unique in that it provides firm-level data, including an objective measure of value-added, and structural firm characteristics. Moreover, individual firms are tracked over a four-year period from 1994/5 to 1997/8. The panel nature of the data allows us to investigate the correlation between firm characteristics and productivity, whilst simultaneously taking into account unobserved firm heterogeneity.

Due to data inconsistencies however, focus is on a sub-two-year panel. Also, some firms reported employee turnover rates well in excess of 1 (the maximum value of turnover rate in the data set is 41!). Since the figure is supposed to measure the turnover of non-causal workers only, the accuracy of these high value responses is questionable. It is suspected that most of those firms that reported a high turnover rate might have mistakenly included the number of newly hired and ceased "casual" employees in their counting. In that case, considerable measurement errors would be introduced. There is no clear pattern on the characteristics of firms with very high reported turnover rates. Thus, observations whose employee turnover rates are greater than 0.8 (equivalent to 5% of total sample) are excluded from the estimations. As the cut-off point of 0.8 is relatively arbitrary, different cut-off points are experimented with as robustness checks.

13.3.2. The empirical model

The empirical model is a productivity function derived from a Cobb–Douglas production function. Using the capital–labour ratio, employee turnover and other firm characteristics to explain productivity, the regression model has the following form:[6]

$$\begin{aligned}\ln(V_{it}/L_{it}) = \beta_0 + \beta_1 \ln(K_{it}/L_{it}) + \beta_2 \ln L_{it} + \delta_1 T_{it} + \delta_2 T_{it}^2 \\ + \mathbf{W}_i'\boldsymbol{\varphi} + \mathbf{Z}_{it}'\boldsymbol{\theta} + u_i + e_{it},\end{aligned} \tag{13.1}$$

where V_{it} is value-added of firm i in year t, and K_{it}, L_{it} and T_{it} denote capital, labour (effective full time employees) and employee turnover rate, respectively. Employee turnover rate is measured by the average of new employees and ceased non-casual employees divided by average non-casual employees at the end of year t and $t-1$. Unobserved firm heterogeneity and idiosyncratic disturbances, are respectively denoted u_i and e_{it}. $\mathbf{W}_i$ is a vector of time invariant firm characteristics, including dummies for family business, incorporation, industry, and firm age and firm size at the first observation year. $\mathbf{Z}_{it}$ denotes a vector of time variant covariates including employment arrangements (ratios of employment on individual contract, unregistered and registered enterprise agreements), other employee related variables (managers to total employees ratio, part-time to total employees ratio, union dummies) and other firm characteristics (innovation status in the previous year, borrowing rate at the end of previous financial year, and export status).

Equation (13.1) can be viewed as a (conditional) productivity-turnover curve (PT).[7] The five scenarios regarding the signs of δ_1 and δ_2 and, thus, the shape of the PT curve and the optimal turnover rate are summarised in Table 13.1.

A priori, one would expect $\delta_1 > 0$ and $\delta_2 < 0$, giving rise to an n-shaped PT curve. This is because, when turnover is very low, job–worker match is unlikely to be optimal as technology and worker characteristics change continuously. Hence, the marginal benefit of increasing the labour market flexibility overwhelms the marginal cost of forgoing some FSHC.

[6] It has been verified that terms with orders higher than two are insignificant. Furthermore, if there are feedback effects of productivity on the turnover rate, one should include lagged terms of T in the equation and/or set up a system of equations. For instance, using U.S. data, Azfar and Danninger (2001) find that employees participating in profit-sharing schemes are less likely to separate from their jobs, facilitating the accumulation of FSHC. However, the short time span of our panel data prohibits us from taking this into account in the empirical analysis.

[7] The effects of turnover on productivity are essentially the same as those on value-added as factor inputs have been controlled for.

Table 13.1.* *Various scenarios of the productivity–turnover curve

Scenario	Shape of PT curve ($T \geqslant 0$)	Interpretation	Optimal turnover rate
$\delta_1 = \delta_2 = 0$	Horizontal	FSHC and job matching effects cancel each other	Undefined
$\delta_1 > 0, \delta_2 < 0$	n-shaped	Job matching effects dominate when T is small, while FSHC effects dominate when T is large	$-\frac{\delta_1}{2\delta_2}$
$\delta_1 < 0, \delta_2 > 0$	U-shaped	FSHC effects dominate when T is small, while job matching effects dominate when T is large	Undefined
$\delta_1 \geqslant 0, \delta_2 \geqslant 0,$ $\delta_1 + \delta_2 \neq 0$	Upward sloping	Job matching effects dominate	Undefined
$\delta_1 \leqslant 0, \delta_2 \leqslant 0,$ $\delta_1 + \delta_2 \neq 0$	Downward sloping	FSHC effects dominate	0

As a result, productivity rises with the turnover rate. Due to the law of diminishing marginal returns, the gain in productivity lessens as turnover increases. Eventually the two effects will net out; further increases in turnover will then lead to a fall in productivity.

In the case of an n-shaped PT curve, the optimal turnover rate is equal to $-0.5(\delta_1/\delta_2)$. The rate is not necessarily optimal from the perspective of firms, as competent employees may leave for a better job opportunity. Neither is it necessarily optimal from the perspective of employees, as there may be involuntary departure. In essence, turnover represents the fact that firms are sorting workers and, reciprocally, workers are sorting firms. As a result, the estimated optimal rate should be interpreted from the production perspective of the economy as a whole. Moreover, the measurement does not take into account the hidden social costs of turnover, such as public expenses on re-training and unemployment benefits, and the searching costs borne by job seekers, and for that matter, hidden social benefits such as higher social mobility.

13.4. Empirical results

13.4.1. Results of production function estimation

Table 13.2 reports the estimation results, on the assumption that the unobserved effects of Equation (13.1) are treated as random, for the base case (the sample with cut-off point of 0.8) as well as for the full sample. A random effects specification is chosen as the estimation is based on a large random sample from the population. Moreover, a fixed effects approach would lead to an enormous loss of degrees of freedom, especially as the data contains only 3 years information (Baltagi, 2001). For the base case, two models are estimated; with and without the restriction of constant returns to scale (CRS). The results indicate that the CRS restriction cannot be rejected, as the coefficient of log labour in the unrestricted model is not significantly different from zero. Accordingly, focus is on the CRS results for the base case in the following discussion (the middle two columns).

The coefficient of log capital is very small. This is not surprising due to the use of non-current assets as a proxy of capital (see Appendix A13 for details). This argument gains support from the negative coefficients of firm age dummies in that the under-estimation of capital is larger for older firms.[8] Since both capital and firm age variables are included as control variables, the mismeasurement of capital should not unduly bias the coefficient of employee turnover.

The coefficient of the ratio of employees on individual contract is significantly positive. This is expected as individual contracts and agreements tend to be more commonly used with more skilled employees, and also because such agreements tend to be used in tandem with performance-based pay incentives. Although it is widely believed that registered enterprise agreements are positively correlated with productivity (Tseng and Wooden, 2001), the results here exhibit the expected sign but the effect is not precisely estimated. Interestingly, productivity is higher for unionised firms and it is particularly significant for those with more than 50 per cent of employees being union members.

The coefficient of the lagged borrowing rate is, as expected, positive, and significant. It is consistent with the theory that the pressure of paying back debts motivates greater efforts in production (Nickell *et al.*, 1992). The manager to total employee ratio appears to have no effect on productivity, while the negative effects of part-time to full-time employee ratio is marginally significant. The latter result is probably due to the fact that part-time workers accumulate less human capital than their full-time counterparts.

The coefficient of innovation in the previous year is insignificant, possibly due to the potentially longer lags involved. Export firms have higher productivity; highly productive businesses are more likely to survive in highly competitive international markets and trade may prompt faster absorption of new foreign technologies. Non-family businesses, on average,

[8] If there is no underestimation of capital stock, other things equal, older firms are likely to have higher productivity due to accumulation of experience.

exhibit 16 per cent higher (labour) productivity than family businesses, whereas incorporated firms are 13 per cent higher than non-incorporated ones. The result signifies the importance of corporate governance, as non-family businesses and incorporated firms are typically subject to tighter scrutiny than their counterparts. Medium and medium large firms have 15 and 20 per cent higher productivity, respectively, than small firms.

13.4.2. Employee turnover and productivity

Focus now turns to the impact of turnover on productivity. The coefficients of employee turnover rate and its square are jointly significant at a 5 per cent significance level, although individually the coefficient of the turnover rate has not been precisely estimated. The two coefficients are positively and negatively signed, respectively, implying an n-shaped PT profile. It indicates that job matching effects dominate when turnover is low, whereas FSHC effects dominate as turnover increases. For the base case, the imputed optimal turnover rate is equal to 0.22.[9] This figure changes very little even if the restriction of constant returns to scale is imposed in estimations.

Although the coefficients of other explanatory variables for the full and trimmed samples are not markedly different, the same is not true of those of turnover rate and turnover rate squared. This indicates that the extremely large turnover rates are likely to be genuine outliers, justifying their exclusion. However, notwithstanding this result, the estimated optimal turnover rates are remarkably stable across samples with different cut-off points (Table 13.3), lying between 0.214 and 0.231, *even though the coefficients are sensitive to the choice of estimation sample*. Firms with a turnover rate higher than 0.5 are likely to be “outliers” as our definition of turnover excluded casual workers.[10] Since the measurement errors are likely to be larger at the top end of the distribution, the effect of employee turnover rate weakens as the cut-off point increases. To balance between minimising the measurement errors on the one hand and retaining sufficient number of observations on the other, the 0.8 cut-off point was chosen as the base case.

Note that despite the coefficient of the turnover rate is individually not significantly different from zero (at 5 per cent) for the base case,

[9] Using 1,000 Bootstrap replications, 93.1 per cent of the replications yielded n-shaped PT curves. The 95 per cent confidence interval for the base case optimal turnover rate is (0.052, 0.334).

[10] As a casual benchmark, policy advisers working for the Australian Government are reported to have very high turnover rates, mainly due to long hours, high stress and lack of a clear career path (Patrick, 2002). Their turnover rate was found to range from 29 per cent to 47 per cent under the Keating government (1991–1996).

Table 13.2. Estimation results from random effect models

	Restrict CRS Full sample		Restrict CRS Turnover < 0.8		Not restrict CRS Turnover < 0.8	
	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.
Log capital–labour ratio	0.189*	0.009	0.188*	0.009	0.184*	0.009
Log labour					0.031	0.023
Turnover rate	−0.016	0.027	0.182	0.113	0.169	0.112
Turnover rate squared	−0.001	0.004	−0.418*	0.182	−0.399*	0.181
Ratio of employment on individual contract	0.131*	0.025	0.133*	0.026	0.128*	0.026
Ratio of employment on unregistered agreement	−0.006	0.031	0.004	0.032	0.001	0.032
Ratio of employment on registered agreement	0.057	0.045	0.062	0.047	0.056	0.047
Ratio of manager to total employment	0.095	0.076	0.098	0.078	0.144#	0.084
Ratio of part-time to total employment	−0.044	0.040	−0.075#	0.041	−0.055	0.043
Union dummy (1–49%)	0.031	0.025	0.026	0.025	0.022	0.026
Union dummy (50%+)	0.086*	0.038	0.082*	0.038	0.077*	0.039
Family business	−0.163*	0.024	−0.164*	0.024	−0.166*	0.025
Incorporated	0.135*	0.026	0.132*	0.027	0.130*	0.027
Export	0.106*	0.023	0.103*	0.024	0.097*	0.024
Innovation ($t-1$)	0.005	0.015	0.000	0.016	0.000	0.016
Borrowing rate ($t-1$)	0.011*	0.005	0.011*	0.005	0.011	0.005
Size: medium	0.154*	0.028	0.153*	0.029	0.116*	0.041
Size: medium–large	0.199*	0.052	0.191*	0.053	0.125#	0.075
Age (less than 2 years)	−0.171*	0.050	−0.171*	0.051	−0.170*	0.052
Age (2 to less than 5 years)	−0.060	0.038	−0.061	0.040	−0.057	0.040
Age (5 to less than 10 years)	−0.017	0.032	−0.014	0.033	−0.013	0.033
Age (10 to less than 20 years)	−0.018	0.030	−0.022	0.031	−0.020	0.032
Constant	3.282*	0.056	3.289*	0.058	3.229*	0.080
Industry dummies	Yes		Yes		Yes	
σ_u	0.472		0.481		0.482	
σ_e	0.301		0.297		0.287	
$\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$	0.711		0.725		0.739	
Number of observations	4472		4249		4249	
Number of firms	2357		2311		2311	
χ^2_{31} test for overall significance	1295.2		1235.0		1194.8	

*Indicate significance at 5% level.

#Indicate significance at 10% level.

Table 13.3. ***Results for robustness checks***

	Turnover rate		Turnover rate squared		Optimal	Sample
	Coef.	Std. err.	Coef.	Std. err.	rate	proportion
1996/97–1997/98						
Turnover < 0.5	0.435	0.175	−1.001	0.422	0.217	0.872
Turnover < 0.6	0.411	0.146	−0.962	0.298	0.214	0.914
Turnover < 0.7	0.178	0.124	−0.385	0.219	0.231	0.938
Turnover < 0.8 (base case)	0.182	0.113	−0.418	0.182	0.218	0.950
Full sample	−0.016	0.027	−0.001	0.004	0	1.0
1995/6–1997/98						
Turnover < 0.8	0.153	0.084	−0.244	0.136	0.313	0.951

which implies a downward sloping PT curve (scenario 5 of Table 13.1), the null hypothesis of an n-shaped PT curve is maintained for three reasons. Firstly, this variable is essentially significant at the 10 per cent level (p-value equals 0.106), or at the 5 per cent level for a one-sided test.[11] Secondly, the optimal turnover rates are very similar across different cut-off points and the coefficients of turnover rate are highly significant for the samples with lower cut-off points than 0.8. This means that the low significance of this variable in the base case is likely to be driven by measurement errors of turnover rates.[12] Finally, the two turnover terms are *jointly* significant, and will necessarily be subject to some degree of collinearity.

The model is also estimated by industry and firm size (with the choices of such being driven by effective sample sizes) and the results are presented in Table 13.4. The retail trade industry has the highest optimal turnover rate of 0.33, compared to 0.24 and 0.22 of the manufacturing and wholesale trade industries, respectively. The retail trade industry also faces the greatest productivity loss from deviating from the optimal rate as it has the steepest PT curve. Figure 13.1 illustrates the PT curve for three different samples (all, manufacturing and small firms). The diagram is a plot of log productivity against turnover rate. The PT curve can be read

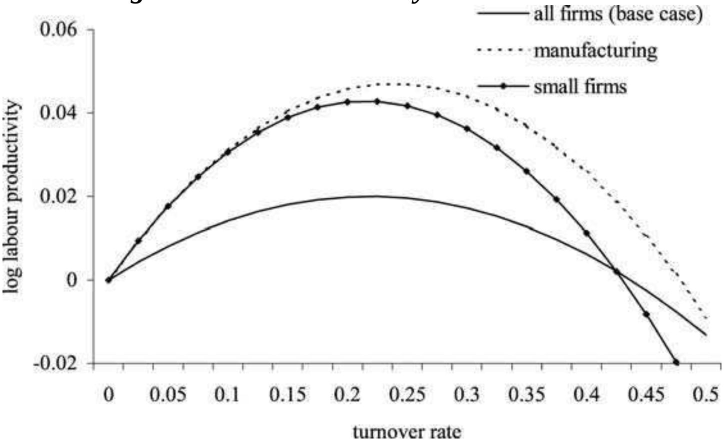
[11] The results presented in this chapter were estimated using STATA 8. The turnover rate variable becomes significant (p-value equals 0.0516) when LIMDEP 8 was used instead, but the magnitude did not change much (coefficient equals 0.185), and the computed optimal turnover rate remained equal to 0.22.

[12] The reason of choosing 0.8 as the cut-off point instead of 0.5, is that this sample yields a more conservative, and realistic, estimate of potential productivity gains, as the lower the cut-off point, the larger are the magnitudes of coefficients. Given similar optimal turnover rates, the productivity gain is the smallest among samples with lower cut-off points.

Table 13.4. Estimation results by industry and firm size

	Turnover rate		Turnover rate squared		Optimal	Number of
	Coef.	Std. err.	Coef.	Std. err.	rate	observations
Manufacturing	0.393	0.140	−0.821	0.226	0.239	1825
Wholesale trade	0.317	0.326	−0.711	0.550	0.223	792
Retail trade	0.834	0.301	−1.251	0.473	0.333	440
Small firms	0.398	0.144	−0.925	0.240	0.215	2082
Medium and medium–large firms	−0.170	0.176	0.254	0.273	–	2167

Figure 13.1. Productivity–turnover curve



as that, in the base case, increasing employee turnover rate from 0 to the optimal point (0.22), on average, raises productivity by 1.95 per cent.

The median turnover rate for the base case sample is 0.14, which is well below the optimal rate.[13] A possible explanation for the large gap between the estimated optimal rate and the sample median is the lack of coordination between agents (employers and employees) in the labour market. For instance, when an employer is pondering whether to layoff an unproductive employee, he/she needs to consider the chance of finding a better replacement within a certain period of time. The chance depends on, amongst other factors, the turnover rates in other firms. Without sufficient information about the employment plan of each other, agents will make

[13] The average turnover rate of the base case sample is 0.183. However, median is a more useful concept here because the average figure is dominated by the high turnover rates of a handful of firms.

changes at a rate lower than what would have been if information were fully revealed. In Appendix B13, a formal model is presented to elaborate this explanation. Another plausible explanation is that there is an enormous amount of friction in the dismissal and hiring process, such as legal restrictions. Yet another possible explanation is that employers may be concerned about non-pecuniary compensation, such as a harmonious working environment, which may or may not sufficiently compensate for inferior job matching. This scenario is likely to be important for small and medium sized firms, which characterise the BLS data.

While the finding cannot pin down exactly what factors attribute to the gap, it indicates how much can be gained by bringing the turnover rate towards the optimal level. The average productivity gain from closing the gap is equal to 1.1 per cent, which is the average increment of productivity for the firms in the base sample if their turnover rates shift from observed to the optimal values, weighted by the firms' value added.[14]

Note that as the analysis in this chapter is based on small and medium firms, it is not possible to draw inferences to the population of all firms. Very large firms typically consist of many sub-units, which could all be considered smaller "firms". Therefore, intra-firm mobility may substitute inter-firm mobility.[15] Also, it is not possible to test the potential long-term effects of turnover on productivity here due to data restrictions. For instance, unfavourable comments on a firm spread by its involuntarily separated employees may damage its corporate image, and thus weaken its attraction to quality potential employees. Therefore, employee turnover may have slightly stronger negative effect in the long run. However, this reputation effect should not be significant for small and medium firms because of their relative size in the labour market. To examine this long run effect (as well as any potential reverse causation effect discussed in footnote 11) requires the use of a longer panel.

13.5. Conclusions

This paper sets out to quantify the impact of employee turnover on productivity. Of the two major theoretical arguments, FSHC theory asserts that

[14] Note that there is the possibility that lower productivity might lead to payroll retrenchment. However, if so, this is likely to have an impact on staffing decisions with lags (for example, due to uncertainty in distinguishing cyclical effects from long run declines in productivity, and measurement error in identifying individual worker's productivity in team production). Since the estimations use contemporaneous turnover and productivity figures, any potential endogeneity will be alleviated.

[15] In a case study, Lazear (1992) finds that the pattern of within-firm turnover from job to job resembles that of between-firm turnover.

high turnover lowers firms' incentives to provide staff training programs and consequently, reduces productivity. On the other hand, job matching theory postulates that turnover can help employers and employees avoid being locked in sub-optimal matches permanently, and therefore increases productivity. The conflict between retaining workforce stability on the one hand, and flexibility on the other, gives rise to the potential existence of an "optimal" turnover rate.

Using an Australian longitudinal data set, productivity was found to be a quadratic function of turnover. The n-shaped PT curve is consistent with the intuition that job matching effects dominate while turnover is low, whereas FSHC effects dominate while turnover is high. The optimal turnover rate is estimated to be about 0.22. This result was robust to both estimation method and sample (with the possible exception of the retail trade sector).

The fact that the estimated optimal rate is much higher than the sample median of 0.14 raises questions about whether there are institutional rigidities hindering resource allocation in the labour market. Using a theoretical model, it is shown that the large turnover gap can be explained by the lack of decision coordination between agents in the market. The empirical results also indicate that higher productivity can be gained from narrowing this gap – average productivity increase was estimated to be at least 1.1 per cent if the turnover rates across the sampled firms are brought to the optimal level.

Appendix A13. The working sample and variable definitions

The first wave of BLS was conducted in 1994/5, with a total effective sample size of 8,745 cases. The selection into the 1995/6 sample was not fully random. Businesses that had been innovative in 1994/95, had exported goods or services in 1994/95, or had increased employment by at least 10 per cent or sales by 25 per cent between 1993/94 and 1994/95, were included in the sample. A random selection was then made on all remaining businesses. These businesses were traced in the surveys of the subsequent two years. In order to maintain the cross-sectional representativeness of each wave, a sample of about 800 businesses were drawn from new businesses each year. The sample size in the second, third and fourth waves are around 5,600. For detailed description of the BLS data set, see Tseng and Wooden (2001). Due to confidentiality considerations, the complete BLS is not released to the public, only the Confidentialised Unit Record File (CURF) is available. In the CURF, businesses exceed 200 employees and another 30 businesses that are regarded as large enterprises using criteria other than employment are excluded. This leaves around 4,200 businesses in the balanced panel.

Deleting observations that had been heavily affected by imputation, as their inclusion would impose artificial stability, further reduced the number of cases available for analysis. Moreover, businesses in the finance and insurance industries were excluded because of substantial differences in the measures of value-added and capital for these firms (and effective sample sizes too small to undertake separate analyses on these groups). In addition, observations with negative sales and negative liabilities were dropped, as were a small number of cases where it was reported that there were no employees. In total, this left just 2,435 businesses in our sample. Summary statistics are presented in Table A13.1.

The dependent and explanatory variables are briefly described as follows:

- $\ln V_{it}$ (log value-added): Value-added is defined as $\textit{sales} - \textit{purchase} + \textit{closing stock} - \textit{opening stock}$, in financial year t.
- $\ln K_{it}$ (log capital): Capital is measured as the total book value of non-current assets plus imputed leasing capital. As reported in Rogers (1999), the importance of leasing capital relative to owned capital varies significantly with firm size and industry, suggesting that leasing capital should be included if we are to accurately approximate the total value of capital employed in the production process. Leasing capital is imputed from data on the estimated value of rent, leasing and hiring expenses.[16]
- $\ln L_{it}$ (log labour): Labour input is measured as the number of full-time equivalent employees.[17] Since employment is a point in time measure, measured at the end of the survey period (the last pay period in June of each year), we use the average numbers of full-time equivalent employees in year t and year $t-1$ for each business as their labour input in year t.[18]

[16] Leasing capital is imputed using the following formula: leasing capital $=$ leasing expenses$/(0.05 + r)$. The depreciation rate of leasing capital is assumed to be 0.05. Ten-year Treasury bond rate is used as the discount rate (r). See Rogers (1999) for more detailed discussion.

[17] The BLS only provides data on the number of full-time and part-time employees while the number of work hours is not available. The full-time equivalent calculation is thus based on estimated average work hours of part-time and full-time employees for the workforce as a whole, as published by the ABS in its monthly Labour Force publication (cat. no. 6203.0).

[18] Capital is also a point in time measure. However, capital is far less variable than labour (especially when measured in terms of its book value), and hence the coefficient of capital is not sensitive to switching between flow and point-in-time measures.

Table A13.1. Summary statistics

	Full sample		Trimmed sample	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
Log labour productivity	4.281	0.695	4.289	0.694
Log capital–labour ratio	3.968	1.091	3.972	1.086
Log labour	2.823	1.119	2.845	1.114
Turnover rate	0.252	0.470	0.183	0.181
Ratio of employment on individual contract	0.251	0.365	0.254	0.367
Ratio of employment on unregistered agreement	0.085	0.249	0.084	0.248
Ratio of employment on registered agreement	0.068	0.218	0.069	0.218
Manager to total employee ratio	0.255	0.169	0.252	0.168
Ratio of part-time to total employee	0.202	0.282	0.195	0.276
Union dummy (1–49%)	0.206	0.405	0.209	0.407
Union dummy (50%+)	0.079	0.270	0.082	0.274
Family business	0.514	0.500	0.512	0.500
Incorporated	0.715	0.451	0.717	0.450
Export	0.271	0.444	0.272	0.445
Innovation $(t-1)$	0.292	0.455	0.293	0.455
Borrowing rate $(t-1)$	0.746	1.395	0.746	1.397
Medium	0.443	0.497	0.445	0.497
Medium–large	0.066	0.248	0.065	0.247
Age (less than 2)	0.062	0.241	0.062	0.241
Age (2 to less than 5 years)	0.129	0.335	0.129	0.335
Age (5 to less than 10 years)	0.248	0.432	0.248	0.432
Age (10 to less than 20 years)	0.288	0.453	0.287	0.453
Age (20 years+)	0.274	0.446	0.275	0.446
Mining	0.008	0.088	0.008	0.088
Manufacturing	0.428	0.495	0.430	0.495
Construction	0.043	0.203	0.042	0.201
Wholesale trade	0.181	0.385	0.186	0.389
Retail trade	0.107	0.309	0.104	0.305
Accommodations, cafes & restaurants	0.036	0.186	0.033	0.180
Transport & storage	0.029	0.169	0.029	0.168
Finance & insurance	0.013	0.113	0.012	0.111
Property & business services	0.118	0.323	0.119	0.324
Cultural & recreational services	0.018	0.133	0.017	0.128
Personal & other services	0.019	0.137	0.019	0.138

- T_{it} (employee turnover rate): Employee turnover rate is measured by the average of new employees and ceased non-casual employees divided by average non-casual employees at the end of year t and $t-1$. The variables are only available from 1995/6 onwards. Moreover, the questions for the calculation of labour turnover rate are slightly different in 1995/6 questionnaires.
- $\mathbf{W}_i$ (time invariant control variables):
  - Firm age dummies: this variable is to control for any bias associated with the mismeasurement of capital, as well as to control for industry specific knowledge.[19]
  - Industry dummies: industry dummies are included to control for industry specific factors that may not be captured by the above variables.
- $\mathbf{Z}_{it}$ (time variant control variables):
  - Employment arrangement: there are three variables included in the regression – proportion of employees covered by individual contracts, by registered enterprise agreements, and by unregistered enterprise agreements. The proportion of employees covered by award only is omitted due to perfect multi-collinearity.
  - Union dummies: these dummies indicate whether a majority or a minority of employees are union members, respectively. A majority is defined as more than 50 per cent and a minority being more than zero but less than 50 per cent. The reference category is businesses without any union members at all.
  - Part-time employee to total employee ratio and manager to total employee ratio: the effect of manager to total employee ratio is ambiguous because a higher ratio implies employees being better monitored on the one hand, while facing more red tape on the other. The effect of part-time to total employee ratio is also ambiguous because part-timers may be more efficient due to shorter work hours, but they may be less productive due to less accumulation of human capital.
  - A dummy variable that indicates whether a business was "innovative" in the previous year: Innovation potentially has a long lag effect on productivity. Since the panel is relatively short, in order to avoid losing observations, we include only a one-year lag. Moreover, the definition of innovation is very board in the BLS. The coefficient of innovation dummy is expected to be less significant than it should be.

[19] A source of measurement bias is the use of the book value of non-current assets. Using the book value will, in general, lead to the underestimation of the true value of capital due to the treatment of depreciation. As firms get older, the book value of capital is generally depreciated at a rate greater than the diminution in the true value of the services provided by the capital stock.

- Dummy variables that indicate whether a business is a family business, or an incorporated enterprise. The questions are asked at the first wave of the survey, so both variables are time invariant.
- Borrowing rate: It is measured at the end of the previous financial year. This variable is used to measure how highly geared a firm is.

Appendix B13. A simple model of optimal turnover rate and coordination

This model is to provide a theoretical explanation for the empirical finding in the main text. The model considers only the coordination problem between firms. We focus on the steady state optimal employee turnover rate for a representative firm. A number of assumptions are in order:

(a) All separations are initiated and controlled by the firm. So there is no employee churning.
(b) Production uses a Cobb–Douglas technology with a fixed capital to labour ratio for both incumbents and newcomers.
(c) The real wages received by both types of worker are fixed.
(d) The degree of job matching is random. As a result, firms are not competing with each other, and all firms benefit from having a larger pool of job seekers.
(e) In every period the firm lays off a certain proportion of incumbents, in the hope of replacing them with better-matched workers.
(f) All incumbents are identical and have the equal chance of being laid off. Therefore, in terms of FSHC, there is a difference between incumbents and newcomers but not amongst incumbents themselves. As a consequence, the output of incumbents depends only on their average tenure but not on the distribution of tenures.

The total number of staff for a representative firm, N, is normalised to one:

$$N = 1 = N_{\mathrm{I}} + N_{\mathrm{H}} - N_{\mathrm{L}}, \tag{B13.1}$$

where N_{I} is the number of incumbents; N_{H} the number of newly hired staff; N_{L} the number of incumbents being laid off in each period. In steady state, the total number of staff remains constant, implying that $N_{\mathrm{H}} = N_{\mathrm{L}}$. So the turnover rate is $\theta = \frac{N_{\mathrm{H}}+N_{\mathrm{L}}}{2N} = N_{\mathrm{H}}$.

Given that the total number of staff is normalised to one and the capital to labour ratio is constant, it implies that the capital stock is fixed. Therefore, the profit of the firm can be written as a function of labour input:

$$\pi = A(N_{\mathrm{I}} - N_{\mathrm{L}})^{\lambda} + B(N_{\mathrm{H}})^{\lambda} - w_{\mathrm{I}}(N_{\mathrm{I}} - N_{\mathrm{L}}) \tag{B13.2}$$

$$- w_{\mathrm{H}}N_{\mathrm{H}} - \frac{c}{2}(N_{\mathrm{H}} + N_{\mathrm{L}}),$$

where A is the productivity factor of incumbents; B the productivity factor of newcomers; w_{I} and w_{H} are the real wage rates for incumbents and newcomers, respectively; $c/2$ the real cost of hiring and laying off staff. Output price is normalised to one.

The amount of FSHC an average incumbent can accumulate is negatively related to the chance that she will be laid off in any given period and, thus, to the turnover rate. Here we specify the productivity factor of incumbents as

$$A = \sigma(1-\theta)^{\alpha}, \tag{B13.3}$$

where σ is a positive coefficient, and its value is positively related to the stock of capital. A larger value of α represents a greater FSHC effect.

The productivity factor of newcomers is not a constant. The firm will try to select candidates with a better job-match than an average incumbent. Otherwise, there would be no gain to lay off experienced staff and find an inexperienced replacement. The average productivity of a newcomer depends on the size of the pool of talent from which firms can pick their candidates. If all firms are identical, then the size of the pool will be positively related to the turnover rate in a representative firm. We specify an ad hoc relationship between them as

$$B = \sigma\theta^{\beta}. \tag{B13.4}$$

The specifications of A and B have the same coefficient σ, because if there are not FSHC and job matching effects, incumbents and freshmen are identical. A larger value of β represents a greater job-matching effect. It is assumed that $\lambda + \alpha < 1$ and $\lambda + \beta < 1$.

If there is no coordination between firms, each firm will treat B as a constant rather than a function of θ. In the following, we consider the two cases that firms do not coordinate and coordinate, respectively.

Without coordination, the problem faced by the firm can be formulated as:

$$\max_{\theta} \pi = \sigma(1-\theta)^{\lambda+\alpha} + B\theta^{\lambda} - w_{\mathrm{I}} - c'\theta, \tag{B13.5}$$

where $c' = c + w_{\mathrm{H}} - w_{\mathrm{I}}$ is the net cost of turnover.

The profit maximising turnover rate $\tilde{\theta}$ is given by

$$(\lambda + \alpha)(1-\tilde{\theta})^{\lambda+\alpha-1} - \lambda\tilde{\theta}^{\lambda+\beta-1} + c'/\sigma = 0. \tag{B13.6}$$

With coordination, the firm treats B as an endogenous variable, and its problem is reformulated as:

$$\max_{\theta} \pi = \sigma(1-\theta)^{\lambda+\alpha} + \sigma\theta^{\lambda+\beta} - w_{\mathrm{I}} - c'\theta. \tag{B13.7}$$

The profit-maximising turnover rate θ^* is given by

$$(\lambda+\alpha)(1-\theta^*)^{\lambda+\alpha-1}-(\lambda+\beta)\theta^{*\lambda+\beta-1}+c'/\sigma=0. \qquad \text{(B13.8)}$$

Using Taylor expansions, it can be shown that $(1-\theta)^{\lambda+\alpha-1}\approx 1+(1-\lambda-\alpha)\theta, \theta^{\lambda+\beta-1}\approx(2-\lambda-\beta)-(1-\lambda-\beta)\theta$. Also, using the fact that all $\tilde{\theta}, \beta$ and $(1-\lambda-\beta)$ are small, it can be stated that $\beta(1-\lambda-\beta)\tilde{\theta}\approx 0$. Applying these to (B13.6) and (B13.8), we can obtain

$$\theta^*-\tilde{\theta}\approx\frac{\beta(2-\lambda-\beta)}{(\lambda+\alpha)(1-\lambda-\alpha)+(\lambda+\beta)(1-\lambda-\beta)}. \qquad \text{(B13.9)}$$

In this equation, λ represents the effect of "pure" labour input, α the effect of FSHC, and β the effect of job matching.

In our empirical study, the sample median is 0.14. This figure corresponds to the case that firms and workers cannot coordinate their decisions, as each individual agent is atomic in the labour market. On the other hand, the estimated optimal turnover rate is about 0.22. This is the figure that a central planner will choose. Therefore, it corresponds to the case that agents can coordinate their decisions. If all turnovers *were* initiated by firms and profit are highly correlated to labour productivity, the empirical finding suggests that $\theta^*-\tilde{\theta}$ is in the order of 0.08 $(=0.22-0.14)$. The value of Equation (B13.9) is much less sensitive to the values of λ and α than to that of β. Thus, we arbitrarily set $\lambda=0.7$ and $\alpha=0.02$. The figures indicate a very small FSHC effect relative to the pure labour effect. As β increases from 0.01 to 0.02 to 0.03, the imputed value of $\theta^*-\tilde{\theta}$ from Equation (B13.9) increases from 0.03 to 0.06 to 0.10. Hence we show that the empirical findings in the main text can be readily explained by just the lack of coordination between firms alone, without even resorting to those between workers and between firms and workers.

References

Abbasi, S.M., Hollman, K.W. (2000), "Turnover: the real bottom line", *Public Personnel Management*, Vol. 29 (3), pp. 333–342.

Ahn, S. (2001), "Firm dynamics and productivity growth: A review of micro evidence from OECD countries", OECD Economic Department Working Papers No. 297.

Azfar, O., Danninger, S. (2001), "Profit-sharing, employment stability, and wage growth", *Industrial and Labor Relations Review*, Vol. 54 (3), pp. 619–630.

Baltagi, B.H. (2001), *Econometric Analysis of Panel Data*, Wiley, New York.

Becker, G.S. (1975), *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, National Bureau of Economic Research, New York.

Borland, J. (1997), "Employee turnover: evidence from a case study", *Australian Bulletin of Labour*, Vol. 23 (2), pp. 118–132.

Burdett, K. (1978), “A theory of employee job search and quit rates”, *American Economic Review*, Vol. 68 (1), pp. 212–220.

Business Asia (1999), “People are assets, too”, September 6, pp. 1–2.

Business Times (2000), “Labor market can absorb retrenched workers, says Fong”, February 11.

Carsten, J.M., Spector, P.E. (1987), “Unemployment, job satisfaction and employee turnover: a meta-analytic test of the Muchinshky model”, *Journal of Applied Psychology*, pp. 374–381.

Cooper, D.P. (2001), “Innovation and reciprocal externalities: information transmission via job mobility”, *Journal of Economic Behavior and Organization*, Vol. 45, pp. 403–425.

Economist (2000), “Employee turnover: labours lost”.

Economist (2001), “Corporate downsizing in America: the jobs challenge”.

Fortune (2002), “The 100 best companies to work for”, 21–39.

Glenn, A., McGarrity, J.P., Weller, J. (2001), “Firm-specific human capital, job matching, and turnover: evidence from major league baseball, 1900–1992”, *Economic Inquiry*, Vol. 39 (1), pp. 86–93.

Iverson, R.D. (1999), “An event history analysis of employee turnover: the case of hospital employees in Australia”, *Human Resource Management Review*, Vol. 9 (4), pp. 397–418.

Jovanovic, B. (1979a), “Job matching and the theory of turnover”, *Journal of Political Economy*, Vol. 87 (5), pp. 972–990.

Jovanovic, B. (1979b), “Firm-specific capital and turnover”, *Journal of Political Economy*, Vol. 87 (6), pp. 1246–1260.

Lazear, E. (1992), “The job as a concept”, in: William, J., Bruns, J., editors, *Performance Measurement, Evaluation, and Incentives*, Harvard Business School, Boston, pp. 183–215.

Lilien, D.M. (1982), “Sectoral shifts and cyclical unemployment”, *Journal of Political Economy*, Vol. 90 (4), pp. 777–793.

Lucas, R.E., Prescott, E.C. (1974), “Equilibrium search and unemployment”, *Journal of Economic Theory*, Vol. 7 (2), pp. 188–209.

Lucifora, C. (1998), “The impact of unions on labour turnover in Italy: evidence from establishment level data”, *International Journal of Industrial Organization*, Vol. 16 (3), pp. 353–376.

Lynch, L.M. (1993), “The economics of youth training in the United States”, *The Economic Journal*, Vol. 103 (420), pp. 1292–1302.

McLaughlin, K.J. (1990), “General productivity growth in a theory of quits and layoffs”, *Journal of Labor Economics*, Vol. 8 (1), pp. 75–98.

Megna, P., Klock, M. (1993), “The impact of intangible capital on Tobin’s q in the semiconductor industry”, *American Economic Review*, Vol. 83 (2), pp. 265–269.

Mok, C., Luk, Y. (1995), “Exit interviews in hotels: making them a more powerful management tool”, *International Journal of Hospitality Management*, Vol. 14 (2), pp. 187–194.

Muchinsky, P.M., Tuttle, M.L. (1979), “Employee turnover: an empirical and methodological assessment”, *Journal of Vocational Behaviour*, Vol. 14, pp. 43–77.

Nickell, S., Wadhwani, S., Wall, M. (1992), “Productivity growth in U.K. companies, 1975–1986”, *European Economic Review*, Vol. 36 (5), pp. 1055–1091.

Patrick, A. (2002), “PM tightens grip on Canberra”, *The Australian Financial Review*, Vol. 6.

Petrongolo, B., Pissarides, C.A. (2001), “Looking into the black box: a survey of the matching function”, *Journal of Economic Literature*, Vol. 39 (2), pp. 390–431.

Rogers, M. (1999), “The performance of small and medium enterprises: an overview using the growth and performance survey”, Melbourne Institute Working Paper 1/99.

Sheehan, E.P. (1993), “The effects of turnover on the productivity of those who stay”, *Journal of Social Psychology,* Vol. 133 (5), p. 699.

Shepard, E.M., Clifton, T.J., Kruse, D. (1996), “Flexible working hours and productivity: some evidence from the pharmaceutical industry”, *Industrial Relationship,* Vol. 35 (1), pp. 123–139.

Tseng, Y.-P., Wooden, M. (2001), “Enterprise bargaining and productivity: evidence from the business longitudinal survey”, Melbourne Institute Working Paper 8/01.

This page intentionally left blank

Panel Data Econometrics
B.H. Baltagi (Editor)
© 2006 Published by Elsevier B.V.
DOI: 10.1016/S0573-8555(06)74014-7

CHAPTER 14

Dynamic Panel Models with Directors' and Officers' Liability Insurance Data

George D. Kaltchev

Department of Economics, Southern Methodist University, 3300 Dyer Street, Suite 301, Dallas,
TX 75275-0496, USA
E-mail address: gkaltche@mail.smu.edu

Abstract

This paper uses a unique US dataset to analyze the demand for Directors' and Officers' liability insurance utilizing dynamic panel models. Some well-established theories propose that corporate insurance plays a role in mitigating agency problems within the corporation such as those between shareholders and managers, and managers and creditors, mitigates bankruptcy risk as well as provides real-services efficiencies. Applying dynamic panel data models, this paper uses these theories to perform empirical tests. The hypothesis that D&O insurance is entirely habit driven is rejected, while some role for persistence is still confirmed. I confirm the real-services efficiencies hypothesis and the role of insurance in mitigating bankruptcy risk. Firms with higher returns appear to demand less insurance. Although alternative monitoring mechanisms over management do not appear to play a large role, I find some support that insurance and governance are complements rather than substitutes. I fail to confirm the role of insurance in mitigating under-investment problems in growth companies.

Keywords: liability insurance, corporate insurance and risk management, shareholder litigation, corporate governance, dynamic panel data models, GMM

JEL classifications: G3, C23

14.1. Introduction

One aspect of corporate finance that has not received much empirical attention is corporate insurance. Mayers and Smith (1982) report that corporations on the aggregate spend more money on purchasing insurance than on paying dividends. Yet that area remains largely unexplored empirically, at least with US data. There is a particular type of insurance that is directly related to corporate governance and the relationship between shareholders and managers. This is Directors' and Officers' (D&O) Liability Insurance, regularly purchased in the US. This insurance plays a significant role in the corporate structure and protects against the risk of managers not fulfilling their contractual obligations towards shareholders and other stakeholders in the company. The present paper uses a unique data set from the US to analyze the demand for D&O Insurance and factors that explain the limits chosen by public companies, thus enriching the relatively small applied literature on these important issues. It examines coverage limits in light of past stock performance, corporate governance and financial risk, using very recent data. The paper manages to confirm some theories on the demand for corporate insurance and tests for the first time corporate insurance theory with US panel data.

The hypotheses to be tested derive from the seminal papers of Mayers and Smith (1982) and MacMinn and Garven (2000). The Main Hypothesis, however, follows from Urtiaga's (2003) game theoretical model. Companies with better returns demand less insurance. The higher the returns, the less likely the shareholders to sue. Moreover, higher returns imply that managers are working in the interest of the shareholders (and creditors) and there are less agency costs. This implicitly supports the agency costs theory; the lower the agency costs, the lower the litigation risk and less insurance is demanded. Thus it is also connected to the theories of Mayers and Smith (1982) and MacMinn and Garven (2000) on the role of insurance in mitigating agency costs. As a result, this hypothesis blends several theories; that is partly the reason it is chosen as main hypothesis. Returns are measured by raw stock returns and returns on assets. It is expected to find a positive correlation of those variables with the limit.

The following are the control hypotheses. Hypothesis 2 is that corporate governance influences the D&O insurance limit. I test whether corporate governance and insurance are substitutes or complements (cf. Holderness, 1990). If they are substitutes, the better the corporate governance of a company, the less insurance is demanded, as the managers are better supervised and less likely to commit misconduct. If governance and insurance are complements, when extending insurance, the insurer encourages or requires the company to better their governance; thus insurance is associated with better corporate governance. Governance is measured by the number of members on the board, percent of insiders and outsiders, CEO/COB (chair of the board) separation, percent blockholdings, number of blockholders, and directors' and officers' ownership (variables are defined in the

Appendix). If governance and insurance are complements, there will be a positive relationship; if they are substitutes there will be a negative relationship. Thus it is not clear whether the expected signs on the governance variables will be negative or positive.

Hypothesis 3 is that companies in financial distress demand more insurance (Mayers and Smith, 1982). The financial situation of companies is measured by financial risk, leverage, volatility. Volatility is measured as in Hull (2000). The financial risk variable is measured as in Boyer (2003). Those variables are hypothesized to be positively correlated with the level of insurance. This hypothesis also implies that smaller companies (in term of asset size) demand more insurance, as they have higher bankruptcy risk.

Hypothesis 4A concerns size. Mayers and Smith (1982) suggest that smaller companies demand more insurance due to real-service efficiencies and proportionately higher bankruptcy costs. Size is this case is measured by ln (assets). The predicted sign is negative. The existence of mergers and acquisitions is expected to increase the insurance limit.

Hypothesis 4B deals with another measure of size: ln(Market Value of Equity). The higher the Market Value of Equity (MVE), the higher the limit, as the higher would be the potential loss. I perceive this as rationality hypothesis: the higher the potential loss, the higher limit is chosen by the managers.

Hypothesis 5 stipulates that corporate insurance alleviates the under-investment problem (between creditors and managers), as shown by MacMinn and Garven (2000). As growth companies are likely to experience more under-investment problems, they are expected to demand more insurance. The variable to test this is growth (market-to-book ratio), defined as

$$\text{Growth (market-to-book)} = \frac{\text{MVE} + \text{Book value of liabilities}}{\text{Book value of total assets}}. \qquad (14.1)$$

This variable measures the growth opportunities of a corporation. The predicted sign is positive.

Hypothesis 6. Consistent with Boyer's (2003) findings, I expect to observe persistence in limits from year to year. The lagged dependent variable is expected to be significant with a positive coefficient. The lagged dependent variable necessitates the use of dynamic panel models.

To sum up, I will interpret Hypothesis 1 as confirming Urtiaga's (2003) model based on returns and the role of insurance in mitigating agency costs, as suggested by Mayers and Smith (1982) and Holderness (1990). Hypothesis 2 tests whether good governance and liability insurance are complements or supplements, as there are competing theories. I will interpret Hypothesis 3 as confirming the bankruptcy risk theory of Mayers and

Smith (1982) and MacMinn and Garven (2000) and the role of insurance in mitigating bankruptcy costs. I will interpret Hypothesis 4A as providing support for the real-services and bankruptcy risk theory and 4B as providing support that managers rationally choose insurance limits based on the potential size of loss. If confirmed, Hypothesis 5 provides support that corporate insurance mitigates the under-investment problem, as stipulated by MacMinn and Garven (2000). Lastly, Hypothesis 6 reveals persistence in the limits. Boyer (2003) interprets this as evidence of habit. Others no doubt will interpret it as evidence of unchanged risk exposure through time and not necessarily of habit.

Table 14.1 defines the variables.

Table 14.1. Variable definitions

Variable	Definition
D&O limits	The amount of insurance coverage the company carries over a period of time (one year)
Limits/MVE	Ratio of limits over market value of equity, limit per value of equity
MVE	Market value of equity
Total assets	Total assets as reported in Compustat
Acquirer	Equal to 1 if company had an acquisition in the past year; 0 otherwise
Divestor	Equal to 1 if company was acquired in the past year; 0 otherwise
Financial risk	−(Book value of assets)/(Book value of liabilities)*(1/volatility)
Leverage	Long term debt/(Long term debt + MVE)
Volatility	Annual volatility prior to insurance purchase based on compounded daily returns
Growth (market-to-book ratio)	(MVE +Book value of liabilities)/(Book value of total assets)
ROA	Return on assets in the year of insurance = Net income (excluding extraordinary items)/Book value of total assets
Raw stock returns	Buy-and-hold raw returns for one year prior to date of insurance purchase
Members	Number of members on the board of directors
Percent of outsiders	Percent of independent directors on the board
Percent of insiders	Percent of directors who are not independent, such as executives, COB, employees or members of their families
CEO = COB	Chief Executive Officer is same as Chair of the Board
D&O ownership	Percent of firm’s shares owned by directors and officers
Number of blockholders	Number of non-affiliated shareholders who hold at least 5% of stock
Percent blockholdings	Percent of company’s stock held by blockholders

The model to be estimated is

$$\ln(\text{Limit})_{i,t} = \alpha \ln(\text{Limit})_{i,t-1} + \beta X_{i,t} + \eta_i + \nu_{i,t}, \tag{14.2}$$

where X includes the variables described above, η are unobserved individual effects, unchanged over time, the $\nu_{i,t}$ are assumed to satisfy $E(\nu_{i,t}) = E(\nu_{i,t}\nu_{i,s}) = 0$ for $t \neq s$. Other standard assumptions are that $\nu_{i,t}$ are uncorrelated with $\text{Limit}_{i,0}$ and the individual effects. The $X_{i,t}$ are allowed to be correlated with η_i. The dependent variable I am trying to explain is the annual amount of insurance purchased (*limit)* and its dependence on the variables mentioned above. The inclusion of a lagged dependent variable makes this a dynamic panel model. It is estimated using Arellano and Bond's (1991) differenced GMM estimator and the system GMM estimator of Blundell and Bond (1998).

14.2. Data and variables

The data set consists of unbalanced panel data for US companies, spanning the years 1997–2003. I have obtained proprietary and confidential data from two insurance brokerages, which consist of about 300 US companies over the years 1997–2003, both private and public. One of them is a leading insurance broker. Since this study focuses on the public companies (and public data are not easily available for private companies), I removed the private companies from the set, which reduces the set to about 180 companies. After removing companies unlisted on Compustat or CRSP, the data set gets reduced to about 150 companies. To use certain panel data techniques, such as fixed effects, I need at least 2 observations per company and for the Arellano–Bond estimation I need at least 3 observations per company. After removing the companies with single observations, the data set reduces to 113 companies. Thus I have 113 companies with insurance data for at least two years and 90 companies with insurance data for at least three years. The sample is small, but such data are not usually publicly available. In addition, there are researchers who apply the Arellano–Bond method on country models, and since the number of countries is finite, their samples are not large. For instance, Esho *et al.* (2004) apply GMM dynamic panel methods with $N = 44$ (and small T).

The data include D&O insurance amounts, quote dates, effective dates, underwriters, SIC codes. The industries of the companies are: Technology (36 companies), Biotechnology and pharmaceuticals (18), Petroleum, mining and agricultural (12), Non-durable goods manufacturing (12), Merchandising (9), Non-banking financial services (6), Durable goods manufacturing (5), Transportation and communications (3), Personal and business services (2), Banking (2), Health services (2), Construction and real estate (1), Newspaper: publishing and print (1), Radio, TV Broadcast (1), Other (3).

This data set allows me to analyze the demand for D&O insurance in the US. Having panel data allows me to study the dynamic decision-making between years regarding corporate insurance, while Core (1997) and O'Sullivan (1997) use cross-section data only. To my knowledge, this is the first study employing US panel data set of D&O insurance data. It is also the first set on which the Arellano–Bond (1991) techniques will be used.

As in Core (1997), it is assumed that officers, directors, shareholders, and insurers have symmetric beliefs about the probability and distribution of D&O losses. The insurer requires seeing the financial statements of the company before extending coverage. Misrepresentation on these financial statements may cause denial of coverage, as the company has misrepresented the risk they pose. This is becoming more common in the 2000's, as insurance companies are more likely to deny coverage after the corporate scandals. The litigation risk is perceived to have increased in the 2000's after the rise of lawsuits and corporate scandals.

14.3. Results

One-step estimation is preferred for coefficient inference (Bond, 2002). The system estimator is more suitable when the number of time-series observations is small, therefore the preferred results are from the one-step system estimations.

The results from the Arellano–Bond estimations of the limit equation are shown in Table 14.2. First of all, the lagged dependent variable is very significant at the 99% level in all estimations in that table, which justifies the use of a dynamic model. We can see the downward bias in the difference GMM estimation, as the coefficient there is much smaller than in the system estimations. Thus it is safe to assume that the coefficient on the lagged limit is .67 and is significant. Last years decision does influence strongly this year's decision on insurance. I achieve a result similar to Boyer's (2003), who also finds a significance of persistence (using an instrumental regression). The theory that persistence is one of the driving forces behind risk management decisions is supported here. In contrast to Boyer (2003), who finds no significance of any other variable, I find some other variables that also influence the decision in the difference and one-step system estimations.

Significant at the 99% level are growth and $\ln(\mathrm{MVE})$ in both the difference and one-step system GMM estimations. The positive coefficient on

Table 14.2. Arellano–Bond and Blundell–Bond dynamic panel-data estimations. Time variable (t): year; significant variable coefficients at 90% or better are in bold

	Difference GMM (one-step)	System GMM (one-step)
Ln(Limit)	Coefficient (Robust Std. Error)	Coefficient (Robust Std. Err.)
Ln(Limit) lagged	**0.267** (.085)	**0.660** (.065)
Growth	**−0.063** (.021)	**−0.054** (.020)
Leverage	**0.653** (.336)	−0.161 (.102)
Risk	−0.000 (.000)	0.000 (.000)
Raw stock returns	−0.010 (.016)	0.003 (.021)
Members	**−0.056** (.03)	−0.03 (.024)
Percent insiders	**0.03** (.017)	−0.003 (.004)
Percent outsiders	**0.039** (.020)	**0.005** (.003)
CEO = COB	0.097 (.085)	0.091 (.079)
D&O ownership	−0.001 (.005)	0.001 (.002)
Percent blockholdings	−0.002 (.003)	0.000 (.004)
Number of blockholders	0.02 (.030)	0.021 (.035)
Ln(assets)	−0.079 (.09)	−0.06 (.076)
Acquirer	−0.017 (.046)	0.01 (.048)
Divestor	0.002 (.068)	−0.032 (.058)
Ln(MVE)	**0.238** (.069)	**0.214** (.062)
ROA	−0.063 (.039)	**−0.097** (.046)
Volatility	**0.106** (.056)	**0.115** (.067)
Year 1999	0.010 (.073)	0.042 (.082)
Year 2000	0.068 (.073)	0.078 (.071)
Year 2001	−0.022 (.072)	0.000 (.073)
Constant	0.029 (.025)	2.66 (1.181)
Sargan test p-value	0.97	
Hansen J test p-value		0.91
M1 p-value	0.01	0.01
M2 p-value	0.76	0.78

Time dummies are included in all estimations. Robust standard errors are robust to heteroskedasticity. Predetermined variables: growth, leverage, risk, raw returns. M1 and M2 are tests for first- and second-order serial correlation. Sargan test (from two-step estimation) and Hansen J tests are tests for over-identifying restrictions.

Market Value of Equity confirms the effect of size and the importance of MVE as a major measure of the size of damages in a potential shareholder lawsuit. Growth appears with a negative but small coefficient. The sign is not as expected. That does not confirm the under-investment Hypothesis 5. Leverage appears with the highest positive coefficient confirming the role of financial distress in choosing limits, but is significant only in the difference GMM estimation. The higher the financial distress, the higher the protection desired. Volatility also shows with a positive coefficient. Members too are only significant in the difference estimation. Members on the board, surprisingly, show with a negative (but small) coefficient. The other variable measuring alternative monitoring mechanisms over management that appears to be significant is percent of outsiders on the board. It is significant in the one-step system GMM estimation and the difference GMM estimation with a small positive coefficient. The more independent the board, the more insurance is demanded, which is more in line with the hypothesis that governance and insurance are complements.

So far the hypotheses that have been confirmed are the Main hypothesis (through ROA, as higher returns lead to less insurance), the hypothesis that independence of the board and D&O insurance are complements, MVE is positively related to insurance limits, and volatility is negatively related to limits. I find no support at all for Hypothesis 5, under-investment, nor for Hypothesis 4A that smaller companies demand less insurance. Persistence is confirmed in this setup.

14.4. Conclusion

This paper provides much needed empirical tests of corporate insurance theory, using recent D&O data from the US. It is the first study to use US panel data and employ dynamic panel data methodology on such data. The methodological contribution is the application of difference and system GMM estimators to D&O insurance data. Given that persistence may be present in different areas of insurance behavior, it may be beneficial to apply these methods in other insurance settings as well.

The Main hypothesis is confirmed: Returns are consistently significant in determining the desired insurance amount. Mostly Returns on Assets, but also Raw Stock Returns, have the expected significant negative effect on limits. Returns are indeed the best signal shareholders have for the performance of managers and a good litigation predictor used by managers. High returns usually indicate that managers are exerting high level of care in the interest of the stakeholders of the company. The presumptions of Urtiaga’s (2003) model receive empirical validation here. Companies in financial distress are shown to demand higher insurance limits. That confirms Mayers and Smith’s (1982) theory as well as the theory of MacMinn and Garven (2000) about the role of insurance in mitigating bankruptcy risk. Indicators of financial health such as leverage and volatility appear to be significant.

Surprisingly, corporate governance does not play a prominent role in the choice of limit. Companies probably do not perceive litigation as a failure of corporate governance but rather as a result of poor performance. Thus the role of corporate insurance in mitigating the agency problems between managers and shareholders, as far as governance provisions are concerned, is dubious. However, I find more evidence that governance mechanisms and insurance are complements rather than substitutes. The growth variable does not show with the anticipated sign, thus I find no confirmation for the theory that this type of insurance mitigates the agency problems between creditors and shareholders. In fact, I consistently reject that theory. While this theory has some theoretical appeal, it received no empirical validation with this dataset.

I find some support for Boyer's (2003) finding of persistence in corporate risk management decisions in this sample. Thus persistence is present both in the US and Canadian data. It is not clear, however, that the significance of the lagged dependent variable can be interpreted as evidence of habit or evidence of unchanged risk exposure. Habit persistence is not the only significant factor, however, as Boyer (2003) has suggested. There is dynamics in risk management decision-making by corporations and one-time observations might be misleading, which underscores the importance of panel data and dynamic models. Companies adjust to changing environments and emphasize considerations that have come to their attention. These mechanisms do not entirely comply with the existing theories for the demand of corporate insurance, but they are not entirely random either.

The results here do not entirely reject the role of persistence but point out a more diverse picture. While persistence plays a role, companies use also some mechanisms to control for risk rooted in insurance theory. Thus corporate risk management serves some useful purposes. Most importantly, the paper finds some confirmation for the theories of Mayers and Smith (1982, 1987), which are considered the cornerstone of modern corporate insurance theory. Also, it illustrates the usefulness of dynamic panel models in this field.

Acknowledgements

I am particularly grateful to the providers of confidential data who made this study possible. I am thankful to participants at the 11th International Conference on Panel Data, an anonymous referee, and the editor.

References

Arellano, M., Bond, S. (1991), "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations", *Review of Economic Studies*, Vol. 58, pp. 277–297.

Blundell, R.W., Bond, S. (1998), "Initial conditions and moment restrictions in dynamic panel data models", *Journal of Econometrics*, Vol. 87, pp. 115–143.

Bond, S. (2002), "Dynamic panel data models: a guide to micro data methods and practice", *Portuguese Economic Journal*, Vol. 1, pp. 141–162.

Boyer, M. (2003), "Is the demand for corporate insurance a habit? Evidence from directors' and officers' insurance", Unnumbered manuscript, HEC-University of Montreal, Canada.

Core, J. (1997), "On the corporate demand for directors' and officers' insurance", *Journal of Risk and Insurance*, Vol. 64, pp. 63–87.

Esho, N., Kirievsky, A., Ward, D., Zurbruegg, R. (2004), "Law and the determinants of property-casualty insurance", *Journal of Risk and Insurance*, Vol. 71, pp. 265–283.

Holderness, C. (1990), "Liability insurers as corporate monitors", *International Review of Law and Economics*, Vol. 10, pp. 115–129.

Hull, J.C. (2000), *Options, Futures and Other Derivatives*, 4th ed., Prentice-Hall, New Jersey. p. 698.

MacMinn, R., Garven, J. (2000), "On corporate insurance", in: Dionne, G., editor, *Handbook of Insurance*, Kluwer Academic Publishers, Norwell, MA, pp. 541–564.

Mayers, D., Smith, C. (1982), "On the corporate demand for insurance", *Journal of Business*, Vol. 55, pp. 281–296.

Mayers, D., Smith, C. (1987), "Corporate insurance and the underinvestment problem", *Journal of Risk and Insurance*, Vol. 54, pp. 45–54.

O'Sullivan, N. (1997), "Insuring the agents: the role of directors' and officers' insurance in corporate governance", *Journal of Risk and Insurance*, Vol. 64, pp. 545–556.

Urtiaga, M. (2003), "An economic analysis of corporate directors' fiduciary duties", *RAND Journal of Economics*, Vol. 34 (3), pp. 516–535.

Panel Data Econometrics
B.H. Baltagi (Editor)

DOI: 10.1016/S0573-8555(06)74015-9

CHAPTER 15

Assessment of the Relationship between Income Inequality and Economic Growth: A Panel Data Analysis of the 32 Federal Entities of Mexico, 1960–2002

Araceli Ortega-Díaz

Tecnológico de Monterrey, Calle del Puente 222, Col. Ejidos de Huipulco, 14380 Tlalpan, México
E-mail address: araceli.ortega@itesm.mx; aortega@sedesol.gob.mx

Abstract

This paper assesses how income inequality influences economic growth by estimating a reduced form growth equation across the 32 Mexican States. Using dynamic panel data analysis, with both urban personal income for grouped data and household income from national surveys, it finds that inequality and growth are positively related. This relationship is stable across variable definitions and data sets, but varies across regions and trade periods.

Keywords: GMM estimator, panel data models, inequality, growth

JEL classifications: CE23, O4, H50

15.1. Introduction

The relationship between economic growth and income distribution is still a controversial topic. When making economic policy, governments are interested in increasing economic growth in order to increase economic welfare. However, economic growth can also lead to an increase in economic inequality, which reduces economic welfare. However, if governments target reductions in income inequality as a way of improving welfare, economic growth may slow, leading again to welfare loss. This dilemma has prompted many researchers to explore the determinants of income inequality, and the channels through which inequality affects economic growth.

On one hand, economic theory suggests that the relation between income inequality and growth differs according to the economic context (market settings). On the other hand, empirical research suggests that divergence in results come from different data quality, period length, omitted variable bias, or even the econometric technique used.

Analysing the results of previous literature, we observe these studies lack a conceptual framework, with which to clearly identify the characteristics of the model we would be interested in analysing under a particular socio-economic scenario, such as the relationship among countries or within a country, developed or underdeveloped countries, perfect or imperfect capital markets, agents' skill level, particular characteristics of economic situation (trade openness, fiscal reforms and others).

Nevertheless, there are a number of important lessons to be learned from the literature. Loury (1981) found that growth and inequality depend on income distribution within and between periods. Thus, an analysis of pure time series or pure cross section would miss mobility and dispersion effects. Moreover under restrictions on borrowing to invest in human/physical capital Galor and Zeira (1993) found that income distribution polarises (into rich and poor), while Banerjee and Newman (1993) found that agents divide into classes with no mobility out of poverty. In both cases, the influence of inequality on growth will depend on initial conditions. Therefore, we should set the country of analysis in a proper economic context before starting drawing conclusions about the relationship between inequality and growth.

The neoclassical standard model of economic growth with technological progress in a closed economy will always predict GDP per capita convergence (Barro and Sala-i-Martin, 1992), where, independently of income distribution within the country, growth can take place. However, Aghion and Williamson (1998) point out since the convergence model assumes perfect capital markets, results may not hold for developing countries. Moreover, Quah (1997) found that assuming each country/state has an egalitarian income distribution, their income dynamics across countries/states may show stratification, persistence or convergence. Such income dynamics, as well as their economic growth may depend on their spatial location, and the countries with which they trade, among other factors. Quah (1997) states that it is not that inequality influences growth or vice versa, but that both have to be analysed simultaneously.[1]

The current work assesses how income inequality influences economic growth across the 32 Federal Entities of Mexico (Mexican States) and

[1] See Forbes (2000), Benabou (1996), Perotti (1996) or Kuznets (1955). Many of the works included in Benabou (1996), suffer from omitted variable bias.

across time. This question is particularly interesting in the case of Mexico, where a rich North contrasts with a backward South, as shown by differences in human capital levels and income distribution, government expenditure, and the level of capital markets imperfections across States.

The originality of our contribution is that it analyses the relation between income inequality and economic growth at the Federal Entity level across time. To my knowledge, this kind of work is the first based in Mexico and contributes to the country case studies on the relationship of income inequality and growth as described by Kanbur (1996), who argues that country case studies rather than cross-country studies will rule income distribution literature over the next two decades.[2]

This paper is organised as follows. In Section 15.2 we set out the model. Section 15.3 explains the data. Sections 15.3 to 15.7 present the estimation of the model as well as some sensitivity analysis. These sections account for the relationship across time, states, and spatial location and reduce the omitted variable bias. Section 15.8 presents our conclusion and possible extensions of the study.

15.2. Model

We examine the influence of income inequality on growth, using a reduced equation like in Forbes (2000) to make our model comparable with those of other studies. We allow for the influence of human capital, dummy variables are introduced for each Mexican State to control for the time-invariant omitted variables bias effect, and time dummies are used to control for aggregate shocks. We estimate Equation (15.1):

$$\begin{aligned}\text{Growth}_{it} = {} & \beta_1 \text{GSP}_{i,t-1} + \beta_2 \text{Inequality}_{i,t-1} \\ & + \beta_3 \text{Human–Capital}_{i,t-1} + \alpha_i + \eta_t + u_{it},\end{aligned} \tag{15.1}$$

where i indexes the states (panel variable) and t is the time variable, α_i are State dummies which can be interpreted as the unobservable State effect, η_i are period dummies denoting unobserved time effects, and u_{it} is the stochastic disturbance term.

[2] Kanbur (1996) points out that while the cross-country literature provides some interesting theories and tests of the development process, its policy implications are not clear. The literature of the process of development and income distribution according to Kanbur has passed through four phases, of which the fourth phase, expected to be found in most of the coming studies, is an intra-country analysis that incorporates the trade off between growth and distribution emphasised in the 1950's (second phase) as well as the short and long run consequences of growth studied in the 1990's (third phase).

15.3. Data sets and measurement

We consider two different data sets. Data set 1 (DS1) covers information from 1960 to 2000, on a decade basis. It considers personal income from grouped data to calculate inequality. Data set 2 (DS2), covers information from 1984 to 2002, on a biannual basis, except for the first two surveys. It considers households income from household surveys to calculate the inequality measures. Sources for the data are listed in Appendix A15.

Schooling is the average year of schooling of the population aged 12 and older.

Literacy is defined as the proportion of the population aged 12 and older who can read and write considered for females and males separately.

Growth is the Gross State Product per capita (GSP) at constant prices.

Income variable is the total after tax cash income received per worker for DS1, and per household for DS2. We did not include non-monetary income because this measured is not specified in all the surveys.

Inequality measure. We use the Gini coefficient to measure inequality because most of the studies choose this measure, and we want to make our results comparable to the results of other surveys. We also use the 20/20 ratio as an alternative measure of inequality and the income share of the third quintile Q3 as a measure of equality.[3]

15.4. Estimation

Following Forbes (2000) and Baltagi (1995), there are three factors considered to estimate Equation (15.1) most accurately: the relation between the State-specific effect and the regressors, the presence of a lagged endogenous variable, and the potential endogeneity of other regressors.

We use dynamic panel data methods to control for the previous problems. The estimation of the model is complex given the presence of a lagged endogenous variable.[4] Considering that GSP_{it} is the logarithm of the per capita Gross State Product for State i at time t, then $\text{growth}_{it} = \text{GSP}_{it} - \text{GSP}_{i,t-1}$, and rewriting Equation (15.1), we get:

$$\begin{aligned} \text{GSP}_{it} - \text{GSP}_{i,t-1} = \beta_1 \text{GSP}_{i,t-1} + \beta_2 \text{Inequality}_{i,t-1} \\ + \beta_3 \text{Schooling}_{i,t-1} + \alpha_i + \eta_t + u_{it}, \end{aligned} \tag{15.2}$$

[3] See Forbes (2000) and Perotti (1996).

[4] It is worth noticing that Barro and Sala-i-Martin (1992) never controlled for this kind of effect.

$$\begin{aligned}\text{GSP}_{it} = {} & \gamma_1 \text{GSP}_{i,t-1} + \beta_2 \text{Inequality}_{i,t-1} + \beta_3 \text{Schooling}_{i,t-1} \\ & + \alpha_i + \eta_t + u_{it},\end{aligned} \tag{15.3}$$

where $\gamma_1 = 1 + \beta_1$.

In matrix notation, it is equivalent to writing:

$$y_{it} = \gamma y_{i,t-1} + X'_{i,t-1} B + \alpha_i + \eta_t + u_{it}, \tag{15.4}$$

$$\begin{aligned} y_{it} - y_{i,t-1} = {} & \gamma (y_{i,t-1} - y_{i,t-2}) + (X'_{i,t-1} - X'_{i,t-2}) B \\ & + (u_{it} - u_{it-1}). \end{aligned} \tag{15.5}$$

Equation (15.5) can now be estimated using the Arellano–Bond (A&B) method. Table 15.1 shows the results of estimating Equation (15.1) with Fixed Effects (FE), Random Effects (RE), and the A&B method using first and second step estimators GMM1 and GMM2 respectively.[5] In each case, we report the results with and without time dummies, and using datasets DS1 and DS2.[6]

The result of the Hausman test shows that the state-specific effects are correlated with the regressors, so the RE estimator is rejected in favour of FE estimator. However, FE is inconsistent (Baltagi, 1995). Thus, the only consistent estimator is the GMM estimator. Therefore, we test for the hypothesis that average autocorrelation in GMM residuals of order 1 and 2 is equal to 0. In general it is not worrying that the 1st-order autocorrelation is violated, it is more important that the 2nd-order autocorrelation is not violated. In our case the GMM with dummies and both GMM2's do not violate that second-order autocorrelation is zero. Finally, from these estimators the only statistically significant for inequality are GMM2 without dummies for DS1 and GMM2 with dummies for DS2.[7] And both coefficients are positive.

As the Arellano–Bond estimator controls for the unobservable time-invariant characteristics of each Mexican state and focuses on changes in these variables within each state across time, the coefficients measure the relationship between changes in inequality and changes in growth

[5] The two-step GMM estimation uses an optimised variance–covariance matrix that corrects the second-order autocorrelation problem. According to Arellano and Bond (1991), "... this apparent gain in precision may reflect downward finite sample bias in the estimates of the two-step standard errors...".

[6] Since the number of years between ENIGH surveys is not the same for the first two surveys as it is for subsequent ones, we perform the analysis dropping 1984 (first survey), and then dropping 1984 and 1989 (second survey). In both cases, the coefficient of inequality is positive and significant.

[7] The significance of dummies in DS2 can be explained by the fact the time between one household survey and another is only two years.

Table 15.1. Growth and inequality regressions using panel methods

Estimation method	FE	RE	FE with year dummies	RE with year dummies	A&B GMM1	A&B GMM1 dummies	A&B GMM2	A&B GMM2 dummies
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Data set 1 (DS1)								
GSP_{t-1}	−0.096**	−0.04**	−0.088**	−0.033**	−0.332**	0.098	−0.488**	0.117**
	(0.011)	(0.007)	(0.010)	(0.006)	(0.131)	(0.157)	(0.075)	(0.025)
Inequality_{t-1}	0.005	−0.019	−0.012	−0.016	0.395**	−0.026	(0.497)**	0.004
	(0.019)	(0.017)	(0.018)	(0.015)	(0.192)	(0.210)	(0.085)	(0.107)
Schooling_{t-1}	0.056**	0.031**	−0.015	0.031**	−0.073	−0.073	−0.193	0.077
	(0.009)	(0.008)	(0.0190)	(0.010)	(0.243)	(0.215)	(0.176)	(0.110)
Dummy 70–80	–	–	0.057**	0.036**	–	–	–	–
			(0.007)	(0.006)				
Dummy 80–90	–	–	0.051**	−0.015*	–	−0.642**	–	0.271**
			(0.013)	(0.007)		(0.102)		(0.030)
Dummy 90–00	–	–	0.080**	0.0004	–	−0.905**	–	−0.051
			(0.020)	(0.009)		(0.151)		(0.032)
R-squared	0.472	0.248	0.720	0.568	–	–	–	–
States	32	32	32	32	32	32	32	32
Observations	128	128	128	128	96	96	96	96
Period	1960–2000	1960–2000	1960–2000	1960–2000	1980–2000	1980–2000	1980–2000	1980–2000

(*continued on next page*)

Table 15.1. (Continued)

Estimation method	FE (1)	RE (2)	FE with year dummies (3)	RE with year dummies (4)	A&B GMM1 (5)	A&B GMM1 dummies (6)	A&B GMM2 (7)	A&B GMM2 dummies (8)
Data set 1 (DS1)								
Hausman test	chi2(3) = 51.50 Prob > chi2 = 0		chi2(6) = 49.75 Prob > chi2 = 0.00		–	–	–	–
Sargan test	–	–	–	–	chi2(5) = 32.18 Prob > chi2 = 0.0	chi2(5) = 8.03 Prob > chi2 = 0.1	chi2(5) = 18.93 Prob > chi2 = 0.0	chi2(5) = 10.5 Prob > chi2 = 0.06
A&B acov res 1st	–	–	–	–	$z = -2.85$	$z = -3.68$	$z = -0.84$	$z = -1.46$
	–	–	–	–	$\Pr > z = 0.004$	$\Pr > z = 0.000$	$\Pr > z = 0.400$	$\Pr > z = 0.145$
A&B acov res 2nd	–	–	–	–	$z = -2.40$	$z = -0.34$	$z = -1.27$	$z = -0.53$
	–	–	–	–	$\Pr > z = 0.016$	$\Pr > z = 0.731$	$\Pr > z = 0.204$	$\Pr > z = 0.596$
DATA SET 2 (DS2)								
GSP_{t-1}	−0.160**	−0.02**	−0.144**	−0.037**	0.255**	0.547***	0.248**	0.532**
	(0.015)	(0.007)	(0.010)	(0.007)	(0.055)	(0.069)	(0.023)	(0.021)
Inequality_{t-1}	0.072**	0.096**	0.003	0.023	0.144**	0.053	0.142***	0.032**
	(0.015)	(0.017)	(0.011)	(0.0144)	(0.037)	(0.040)	(0.009)	(0.014)
Schooling_{t-1}	0.036	0.095**	0.026	0.081**	0.046	0.030	0.076	0.066
	(0.046)	(0.038)	(0.030)	(0.032)	(0.103)	(0.101)	(0.042)	(0.036)
R-squared	0.470	0.126	0.800	0.4564	–	–	–	–

(*continued on next page*)

Table 15.1. (Continued)

Estimation method	FE	RE	FE with year dummies	RE with year dummies	A&B GMM1	A&B GMM1 dummies	A&B GMM2	A&B GMM2 dummies
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Data set 1 (DS1)								
States	32	32	32	32	32	32	32	32
Obs	224	224	224	224	192	192	192	192
Period	1984–2002	1984–2002	1984–2002	1984–2002	1989–2002	1989–2002	1989–2002	1989–2002
Hausman test	chi2(4) = 109.99 Prob > chi 2 = 0.000		chi2(11) = 203.50 Prob > chi = 0.000		–	–	–	–
Sargan test	–	–	–	–	chi2(20) = 106 Pr > chi2 = 0.0	chi2(20) = 18 Pr > chi2 = 0.58	chi2(20) = 30 P > chi2 = 0.06	chi2(20) = 24 P > chi2 = 0.21
A&B acov res 1st	–	–	–	–	$z = -4.99$ Pr > z = 0.00	$z = -3.63$ Pr > z = 0.00	$z = -4.20$ Pr > z = 0.00	$z = -4.07$ Pr > z = 0.00
A&B acov res 2nd	–	–	–	–	$z = -3.12$ Pr > z = 0.00	$z = 0.60$ Pr > z = 0.54	$z = -3.27$ Pr > z = 0.00	$z = 0.82$ Pr > z = 0.41

Note: The dependent variable is average annual per capita growth. Standard errors are in parentheses. R-squared is the within R-squared for the fixed effects (FE) model and the overall R-squared for random effects (RE). A&B acov res 1st and 2nd is the Arellano–Bond test that average autocovariance in residuals of order 1 and 2, respectively is 0.

*stands for significance at 5%.

**stands for significance at 1%.

***stands for significance at 10%.

within a given state (see Forbes, 2000). This result implies that in the short run (considering periods of ten years each for DS1 and two year periods for DS2) positive changes in lagged inequality are associated with positive changes in natural log GSP (i.e. current GSP growth) within each state across periods. This is in contradiction with both political economy models (Alesina and Rodrik, 1994) and with the models that stress capital market imperfections (Galor and Zeira, 1993; Banerjee and Newman, 1993).

In the following sections we address the following questions: is it only the method of estimation that makes the relation between growth and inequality differ from other results? How robust is this relationship?

15.5. Factors that might affect the coefficient of inequality

Factors such as data quality, outliers, period coverage, and method of estimation might affect the coefficient of inequality; as well as different definitions of inequality and literacy. In this section we check if any of these factors have an impact on the inequality coefficient using the valid A&B estimator.

15.5.1. Data quality

We estimate Equation (15.5) using an alternative source for the per capita Gross State Product that comes from Esquivel (1999) for DS1. The results show the same sign for the coefficient of inequality as before, but the coefficient is not significant. What is important is that for the benchmark estimations in Table 15.1, changes in inequality are positively related to changes in growth and that the data source does not affect the sign of the coefficient.[8]

15.5.2. Outliers

There are three states with different behaviour compared to the 29 remaining states; these are Campeche and Tabasco, which are oil producers, and Chiapas, which is a very poor state. They have been treated differently in the literature, as in Esquivel (1999). When we control for outliers, the sign on inequality does not change, but significance slightly increases.

[8] Due to space problems we do not report all the estimations, but they are available from author on request.

15.5.3. Periods coverage and method of estimation

A third factor that may affect the coefficient is the length of the periods considered, so we performed several estimations of Equation (15.1) varying the period lengths. First, for each data set, we consider one long period (1960 to 2000 for DS1, and 1984 to 2002 for DS2), as the long-term period, then Equation (15.1) has to be rewritten as Equation (15.6), and be estimated for one long period with OLS.

$$\text{Growth}_i = \alpha_0 + \beta_1 \text{Income}_i + \beta_2 \text{Inequaltiy}_i + \beta_3 \text{Schooling}_i + u_i. \tag{15.6}$$

The problem with Equation (15.6) is that it suffers from bias caused by the endogenous lagged variable, and due to the few observations available for this type of specification, it is better to consider other type of specification. Hence, we divide the 40-year period for DS1 into three short periods according to the degree of trade openness. We consider the period before Mexico joined the GATT (1960–1980) as the Non-Trade period (although trade was taking place), then we consider the GATT period as the period between joining GATT and before signing NAFTA (1980–1990). The last period will be the NAFTA period (1990–2000).[9]

Then, still using these three short periods, we use A&B estimator with trade period dummies for DS1, using Equation (15.4). The inequality estimate is negative with GMM1 and positive with GMM2 but is not significant (see Table 15.2). In both cases the dummies have a negative sign and are statistically significant.

Finally, we divide the 18-year period for DS2 into two short periods according to the degree of trade openness, the GATT period (1984–1994) and the NAFTA period (1994–2002). Again we estimate Equation (15.4) using A&B estimator, separately for each trade period. The results are given in Table 15.2. The two periods show a positive and significant coefficient. Finally, we use all periods, but adding a dummy for GATT period, and then for NAFTA period. The inequality estimate is positive and very significant. These estimations have the same coefficients except for the sign in the dummy variable: when we include the GATT dummy it is positive and significant, but the opposite is found when we include the NAFTA

[9] According to Boltvinik (1999), in the period 1983–1988, the fight against poverty and inequality was discontinued. New efforts and programs started in the 1988 presidential period, including Solidaridad and Progresa (nowadays Oportunidades). On the other hand, economic policies for the period before signing of the GATT were based on import substitution and expenditure-led growth, but after signing, an export-led growth policy was implemented (e.g., Székely, 1995). Putting together these facts may explain why before 1988 the relationship is negative and afterwards positive.

Table 15.2. Effects of varying period length and estimation method

	Data set DS1		Data set DS2					
Method	A&B GMM1 period of trade	A&B GMM2 period of trade	A&B GMM1 GATT	A&B GMM2 GATT	A&B GMM1 NAFTA	A&B GMM2 NAFTA	A&B GMM1 ALL	A&B GMM1 ALL
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GSP_{t-1}	0.097	0.117**	0.381**	0.451**	−0.185	−0.211**	0.235**	0.235**
	(0.157)	(0.025)	(0.096)	(0.077)	(0.139)	(0.092)	(0.052)	(0.052)
$Ineq_{t-1}$	−0.026	0.004	0.151**	0.121**	0.121**	0.121**	0.090**	0.090**
	(0.210)	(0.107)	(0.069)	(0.051)	(0.042)	(0.024)	(0.037)	0.037
$Scho_{t-1}$	−0.074	0.077	0.135	0.201*	−0.091	−0.140	0.008	0.008
	(0.215)	(0.110)	(0.144)	(0.116)	(0.148)	(0.154)	(0.097)	0.097
Dummy	−0.642**	−0.593**	–	–	–	–	0.052**	
GATT	(0.102)	(0.039)					(0.013)	
Dummy	−0.905**	−0.813**	–	–	–	–	–	−0.052**
NAFTA	(0.151)	(0.092)						0.013
States	32	32	32	32	32	32	32	32
Periods	3	3	2	2	4	4	6	6
Years	1960–2000	1960–2000	1984–1994	1984–1994	1994–2002	1994–2002	1984–2002	1984–2002
Sargan test	chi2(5) = 8.0	chi(5) = 10	chi(20) = 2	chi(20) = 3	chi(20) = 43	chi(20) = 23	chi(20) = 105	chi(20) = 105
	Pr > chi = 0.1	Pr > chi2 = 0	Pr > chi2 = 1	Pr > chi2 = 1	Pr > chi2 = 0	Pr > chi2 = 0.25	Pr > chi2 = 0	Pr > chi2 = 0
A&B acov	$z = -3.68$	$z = -1.46$	$z = -2.55$	$z = -3.17$	$z = -1.64$	$z = -1.39$	$z = -5.09$	$z = -5.09$
res 1st	Pr > z = 0.000	Pr > z = 0.14	Pr > z = 0.01	Pr > z = 0.00	Pr > z = 0.10	Pr > z = 0.16	Pr > z = 0.0	Pr > z = 0.0
A&B acov	$z = -0.34$	$z = -0.53$	–	–	$z = -2.76$	$z = -2.11$	$z = -1.27$	$z = -1.27$
res 2nd	Pr > z = 0.73	Pr > z = 0.59			Pr > z = 0.005	Pr > z = 0.03	Pr > z = 0.20	Pr > z = 0.20

Notes: Dependent variable is average annual per capita growth.
*stands for significance at 5%.
**stands for significance at 1%.

dummy. The change in sign across periods suggests that the relationship between inequality and growth has been changing over time. One of the reasons for this change might be trade openness.[10]

We conclude that time length and the period studied may affect the relation between inequality and growth. The NAFTA period is difficult to interpret as its initial stage coincides with the Mexican economic crisis in December 1994.[11]

15.5.4. Different definitions of inequality and literacy

In this section we analyse whether changing the human capital variable from schooling to literacy has any effects. We find that changing the human capital variable only affects the sign of the inequality coefficient for the last trade period. The rest of the inequality coefficients remain the same in sign and significance.

15.6. Grouping and regional analysis

In this section we examine the idea of Quah (1997) about club formation: that rich states are located near rich states and poor near poor ones. We are interested in testing whether the clubs have different relationships between inequality and growth. We group the States using different methods.

We first use the method of Esquivel (1999) to group the States according to location to see if there is any difference in the inequality regression coefficient across regions, as we can find intrinsic characteristics that make economies differ.[12] The results in Table 15.3 show that the inequality coefficient is positive in 71% of the cases, but only significant in 43% of them, probably due to the small number of observations within each group (*NT* is too small).

[10] Barro's estimations, described in Banerjee and Duflo (1999), and which describe a *U* shape relationship between growth and inequality during 1960–1995 for poor countries, and positive for Latin-America, do not contradict the signs obtained by our three period estimation $(-, -, +)$.

[11] Székely (1995) argues that it is still early to judge the new economic model that currently rules economic decision-making in Mexico and which consists mainly of trade liberalisation. Perhaps when the government implements a policy to lessen inequality, financed by an increase in taxes, inequality decreases but growth does also, because incentives for savings decrease see Perotti (1996). This may explain why we find a positive relationship in the NAFTA period.

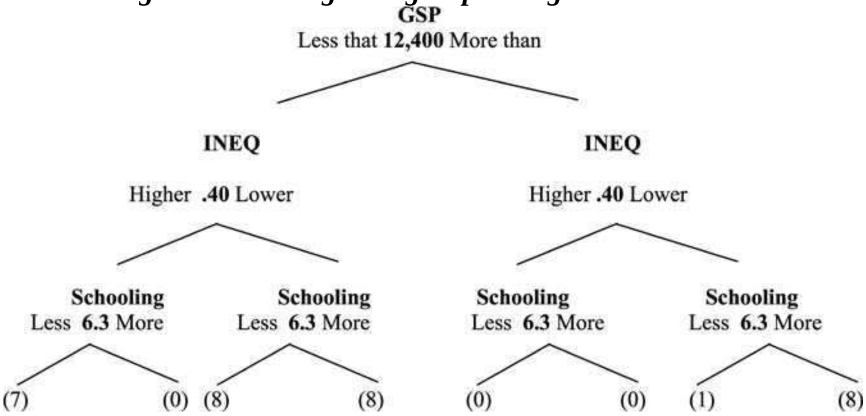
[12] The North for instance, closest to USA, has six of the States with the highest product per capita, and the highest share of foreign direct investment. In contrast, the 57% of the indigenous population is concentrated in the southern regions, its average years of schooling is 5.7 and 6.7 years, compared with 9.6 in D.F., and has poor access to public services.

Table 15.3. *Effect of regional differences on the inequality coefficient*

Geographical regions	Coefficient on INEQ	Standard error	States	Obs	Coefficient on INEQ	Standard error	States	Obs
	Data set DS1 for 1980–2000				Data set DS2 for 1989–2002			
Esquivel definition								
North (0.322)	−1.163**	0.276	6	24	0.097	0.079	6	36
Capital (0.329)	0.814**	0.175	2	8	0.213	0.221	2	12
C. North (0.337)	0.454**	0.204	6	24	0.088	0.068	6	36
Golf (0.337)	−0.449	1.430	5	20	0.037	0.136	5	30
Pacific (0.344)	0.660	0.675	5	20	0.119*	0.057	5	30
South (0.378)	0.783	0.989	4	16	0.149*	0.075	4	24
Centre (0.398)	0.574**	0.192	4	16	0.111	0.084	4	24
Tree definition								
R1	0.552	0.868	7	21	0.128	0.069	7	42
R2	0.572**	0.221	8	24	0.120*	0.057	8	48
R3	0.410	0.270	8	24	0.075	0.063	8	48
R5	−0.381	0.339	8	24	0.131	0.070**	8	48
INEGI’s definition								
W1	0.909	1.168	3	9	0.154	0.094	3	18
W2	0.455	0.533	6	18	0.207**	0.074	6	36
W3	−0.873	0.685	3	9	0.041	0.074	3	18
W4	0.656**	0.234	9	27	0.027	0.063	9	54
W6	−0.353	0.241	9	27	0.092	0.061	9	54

Note: Initial Gini coefficient is in brackets, showing geographical regions are ranked by initial inequality (ascendant).
*stands for significance at 5%.
**stands for significance at 1%.

Figure 15.1. Regional groups using a tree structure



Next, we re-estimate the model grouping states inspired on the tree algorithm technique used in Durlauf and Johnson (1995), but without optimisation. The tree technique consists in splitting the 32 States into two groups, according to their GSP. Afterwards each group is split into two according to their level of inequality. Finally, each of the four groups is split according to their schooling level. With this technique, we have five groups, which we can use to define our own welfare regions, where region 1 has the lowest welfare and region 5 the highest welfare. The resulting tree is shown in Figure 15.1.

The results in Table 15.3 show that the richest region (the ones in the right part of the tree that enjoy the highest GSP, highest schooling level and lowest inequality) has a negative sign on inequality coefficient. The rest of the regions have a positive coefficient. However, results are still not significant, so we cannot derive a strong conclusion from these results.

The National Statistics Office in Mexico (INEGI) performs a welfare analysis where it divides the Federal Entities according to their level of well-being which takes into account 47 socio-economic variables like population characteristics and infrastructure. They use cluster analysis and group the Federal Entities in seven groups, where the lowest level of welfare is rated as level one, to the highest level of welfare that is rated as seven. The estimation results using this information (in Table 15.3) show the same pattern as before, the richest region has a negative coefficient but results are significant only in 20% percent of the cases. Using DS2 instead of DS1, all coefficients on inequality become positive, but significance is still a problem.

Since economic performance and income are highly related, we divide our data according to their income level. We do this by considering the interval defined by the minimum and maximum GSP levels across the 32

Table 15.4. Regression results according to initial GSP groups

	Initial GSP groups	Coefficient on INEQ	Standard error	States	Obs	Period of growth
Data set DS1						
Using INEGI data						
	Poor < 6037	0.506**	0.285	17	51	1980–2000
	$6037 \leqslant$ Mid < 9932	−0.048	0.323	10	30	
	Rich $\geqslant 9932$	0.225	0.213	5	15	
Using G. Esquivel data						
	Poor < 9000	0.185	0.265	17	51	1980–2000
	$9000 \leqslant$ Mid < 16000	0.067	0.480	10	30	
	Rich $\geqslant 16000$	−0.111	0.467	5	15	
Data set DS2						
Initial GSP groups						
	Poor < 13330	0.088**	0.039	17	102	1989–2002
	$13330 \geqslant$ Mid < 19800	0.124*	0.069	11	66	
	Rich $\geqslant 19800$	0.212	0.132	4	24	

Note: States are categorised based on GSP per capita in 1990. Income is measured in 1993 pesos.
*stands for significance at 5%.
**stands for significance at 1%.

Federal Entities. We split this interval in three equal parts and define as "poor" those States whose income fall into the lowest part of interval, "mid" those whose income fall in the mid interval, and "rich" those with income in the top interval.

In Table 15.4, we can observe that the group of the poorest States has a positive coefficient on inequality, but the level of significance changes. The coefficient of the middle and richest States varies as well as its significance. But results are in line with those observed in the grouped data by regions in the previous sections.

We can conclude from Section 15.6 that the relationship between inequality and growth shows a strong contrast between poor and rich regions, northern and southern regions. For rich regions (northern) inequality seems to have a negative coefficient. However, for the poorest regions, inequality's coefficient is positive.

Table 15.5. Different inequality measures

Data set	Inequality definitions	Coefficient on INEQ	Standard error	States	Obs	Period of growth
DS1	20/20 Ratio	0.175**	0.070	32	96	1980–2000
	20/20 Ratio no oil	0.223**	0.057	30	90	1980–2000
	POVCAL	0.349	0.247	32	96	1980–2000
	POVCAL No oil	0.578**	0.181	30	90	1980–2000
	Q3	−0.143**	0.025	32	96	1980–2000
DS2	Inequality definitions					
	20/20 Ratio	0.056**	0.014	32	192	1989–2002
	Q3	−0.094**	0.008	32	192	1989–2002

**stands for significance at 1%.

15.7. Analysis with different inequality measures

Finally, recent literature argues that the relationship between income inequality and growth might depend on the definition of the GINI coefficient. Thus, we swap the Gini coefficient calculated with the Yitzhaki–Lerman formula (see Chotikapanich and Griffiths, 2001) with the Gini calculated with the POVCAL formula developed by Chen, Datt and Ravallion. Afterwards, we use the 20/20 ratio as an alternative measure of inequality. The 20/20 ratio is the quotient between the income of the twenty percent of the richest population and the 20 percent of the poorest. Finally, we use Q3, which is the share of income held by the middle quintile. Perotti (1996) uses Q3 as a measure of equality, and Forbes add a negative sign to Q3 to use it as measure of inequality. We will follow Perotti (1996).

Results are shown in Table 15.5. The estimated inequality coefficient is still positive and very significant in all cases. When we use the equality measure Q3, the estimated equality coefficient becomes negative. These results confirm the robustness of the positive relationship between inequality and growth.

15.8. Conclusions and possible extensions

Results coming from this work have to be treated with reasonable caution due to the limited amount of data used. Using two different data sets to account for the influence that the source may have on the results, and using dynamic panel data methods to control for possible omitted variable bias on the estimates, and the endogeneity of the lagged variable, we have found that the relationship between income inequality and economic growth is positive. This result is robust to the use of different measures of per capita Gross State Product, of human capital variable definitions, and measures of inequality. This implies that the data source and variable measures do not affect the sign in our estimation.

We also analyse the impact that varying the period length and the method of estimation has on the sign of the income inequality coefficient. We found that the inequality coefficient is positive and significant when we use DS2, and negative but not significant using DS1. Including a dummy for GATT and NAFTA periods, with DS2 suggest that NAFTA has a negative influence on inequality whereas GATT had a positive influence. This finding could be interpreted meaning that as the Mexican economy becomes more open, the relation between growth and inequality is changing over time. Our results show that time length and the period studied affects the relationship between inequality and growth. Using different grouping methods to test whether club formation affects the coefficient of inequality, we found that the coefficient of inequality is positive for the poorest regions, and tend to be negative for the richest regions. Nevertheless, we cannot draw a conclusion since we lack a sufficient number of observations in each group.

The results from the dynamic panel data estimations suggest that changes in income inequality and changes in economic growth, from 1960 to 2000 and from 1984 to 2002 across the 32 Federal Entities of Mexico, are positively related. This may suggest that high income-inequality is beneficial for growth in that it can stimulate capital accumulation (Aghion and Williamson, 1998).

Further research is needed using different measures adjusted for household needs, in order to explore the robustness of the relationship between inequality and growth. However, we are not only interested in testing the robustness of the sign, but in analysing the channels through which inequality influences growth, using structural equations, as well as in performing a complementary analysis with growth accounting factors, sources of growth and determinants of income inequality.

Acknowledgements

I would like to thank the participants of the following conferences; the Fourth Summer School: Polarization and Conflict/XX Summer Courses/ XIII European Courses in Basque Country University; the IESG 26th Annual Conference; the 1st Mediterranean Summer School in Theoretical and Applied Economics; the UNU-WIDER seminar; the 8th LACEA Annual Meeting; the Weekly Seminar at the CEE, El Colegio de México; the

11th International Panel Data Conference; and the 2004 LAMES Meeting, for their excellent comments. I thank Professor Tony Shorrocks for his econometric comments and to Professor Stephen Jenkins for reading the chapter and giving me excellent feedback, and to my supervisors at the University of Essex Professor V. Bhaskar and Dr. Gordon Kempt; as well as to CONACYT for its financial support. Any remaining mistakes are my own.

Appendix A15. Summary statistics

Table A15.1.

Variable	Definition	Source	Year	Mean	Std. Dev.	Min	Max
DATA SET 1 (DS1)							
Schooling	Average years of schooling of the population	SEP	1960	2.46	0.91	1.00	5.00
			1970	3.19	0.89	1.80	5.80
			1980	4.31	0.95	2.50	7.00
			1990	6.29	1.00	4.20	8.80
			2000	7.53	1.00	5.70	10.20
GSP	Ln of Real GSP per capita in 1993 pesos. Correcting with national deflator before 1990	INEGI	1960	8.60	0.47	7.60	9.46
			1970	8.75	0.38	7.93	9.60
			1980	9.29	0.39	8.56	10.40
			1990	9.23	0.41	8.53	10.16
			2000	9.49	0.43	8.71	10.56
Inequality	Inequality measured by the Gini Coefficient using Leman and Yitzhaki formula. Considering monetary persons income	SE (1960), SIC (1970) SPP (1980) INEGI (2000)	1960	0.38	0.05	0.20	0.47
			1970	0.43	0.06	0.32	0.57
			1980	0.45	0.03	0.40	0.54
			1990	0.37	0.02	0.34	0.48
			2000	0.41	0.03	0.34	0.51
Female literacy	Share of the female population aged over 15 (10) who can read and write	INEGI	1960	63.99	15.83	34.93	85.58
			1970	73.22	12.05	50.38	87.92
			1980	79.61	10.91	54.94	92.28
			1990	85.09	8.74	62.35	94.52
			2000	88.70	6.90	69.95	96.08
Male literacy	Share of the male population aged over 15 (10) who can read and write	INEGI	1960	70.84	12.26	44.87	92.17
			1970	79.06	8.75	59.66	94.31
			1980	85.52	7.21	68.94	96.89
			1990	89.95	5.15	77.52	97.87
			2000	92.11	4.02	82.86	98.26

(*continued on next page*)

Table A15.1. (Continued)

Variable	Definition	Source	Year	Mean	Std. Dev.	Min	Max
DATA SET 1 (DS1)							
GSP2	Ln of Real GSP per capita In 1995 pesos. Correcting for 2000	G. Esquivel	1960	9.07	0.44	8.32	10.05
			1970	9.46	0.46	8.56	10.38
			1980	9.77	0.43	8.95	10.65
			1990	9.77	0.44	9.04	10.84
			2000	9.79	0.41	9.01	10.80
DATA SET 2 (DS2)							
GSP	Ln of Real Gross State Product (GSP) per capita in 1993 pesos. Correcting with national deflator before 1990. Calculating 2002 using national GDP 2002 and State's share in 2001	INEGI	1984	2.56	0.42	1.92	4.01
			1989	2.43	0.42	1.72	3.37
			1992	2.44	0.41	1.77	3.43
			1994	2.50	0.42	1.82	3.53
			1996	2.46	0.42	1.78	3.47
			1998	2.49	0.42	1.77	3.55
			2000	2.58	0.43	1.84	3.66
			2002	2.54	0.43	1.83	3.64
Inequality	Inequality measured by the Gini Coefficient of monetary household income	ENIGH	1984	0.42	0.05	0.27	0.52
			1989	0.47	0.06	0.34	0.63
			1992	0.55	0.06	0.43	0.72
			1994	0.47	0.05	0.37	0.60
			1996	0.49	0.05	0.42	0.71
			1998	0.51	0.04	0.41	0.61
			2000	0.50	0.05	0.37	0.58
			2002	0.47	0.04	0.37	0.56
Female literacy	Share of the female population aged over 15 (10) who can read and write	ENIGH	1984	84.54	9.78	64.38	98.31
			1989	85.65	8.53	62.73	97.03
			1992	82.37	9.87	60.92	94.90
			1994	83.12	9.37	60.05	94.77
			1996	84.66	7.65	64.84	95.21
			1998	85.56	7.93	69.55	97.90
			2000	86.88	5.98	73.75	95.59
			2002	87.18	7.63	70.26	97.10
Male literacy	Share of the male population aged over 15 (10) who can read and write	ENIGH	1984	91.05	6.23	79.38	100.00
			1989	88.71	5.84	78.72	97.05
			1992	86.13	6.47	73.97	97.66
			1994	86.93	6.05	70.83	97.67
			1996	88.17	4.63	76.57	97.37
			1998	87.46	5.65	73.09	97.53
			2000	88.49	4.37	81.08	98.37
			2002	89.14	5.77	75.00	97.67

References

Aghion, P., Williamson, J. (1998), *Growth, Inequality and Globalization*, Cambridge University Press.

Alesina, A., Rodrik, D. (1994), “Distributive politics and economic growth”, *Quarterly Journal of Economics*, Vol. 109 (2), pp. 465–490.

Arellano, M., Bond, S. (1991), “Some test of specification for panel data: Monte Carlo evidence and an application to employment equations”, *Review of Economic Studies*, Vol. 58, pp. 277–297.

Baltagi, B.H. (1995), *Econometric Analysis of Panel Data*, John Wiley, Chichester.

Banerjee, A., Duflo, E. (1999), “Inequality and growth: what can the data say?”, Mimeo, MIT, pp. 1–32.

Banerjee, A., Newman, A. (1993), “Occupational choice and the process of development”, *Journal of Political Economy*, Vol. 101 (2), pp. 274–298.

Barro, R., Sala-i-Martin, X. (1992), “Convergence”, *Journal of Political Economy*, Vol. 100 (2), pp. 223–251.

Benabou, R. (1996), “Inequality and growth”, NBER, pp. 11–72.

Boltvinik, J., Hernández, L. (1999), *Pobreza y Distribución del Ingreso en México*, Siglo Veintiuno Editores.

Chotikapanich, D., Griffiths, W. (2001), “On calculation of the extended Gini coefficient”, *Review of Income and Wealth*, Vol. 47 (1), pp. 541–547.

Durlauf, S., Johnson, P. (1995), “Multiple regimes and cross-country growth behaviour”, *Journal of Applied Econometrics*, Vol. 10 (4), pp. 365–384.

Esquivel, G. (1999), “Convergencia regional en México, 1940–1995”, Cuaderno de trabajo de El Colegio de México, no. IX-99.

Forbes, K. (2000), “A reassessment of the relationship between inequality and growth”, *American Economic Review*, Vol. 90 (4), pp. 869–887.

Galor, O., Zeira, J. (1993), “Income distribution and macroeconomics”, *Review of Economic Studies*, Vol. 60, pp. 35–52.

Kanbur, R. (1996), “Income distribution and development”, in: *Handbook on Income Distribution*, North-Holland, pp. 1–41.

Kuznets, S. (1955), “Economic growth and income inequality”, *American Economic Review*, Vol. 45, pp. 1–28.

Loury, G. (1981), “Intergenerational transfers and the distribution of earnings”, *Econometrica*, Vol. 49 (4), pp. 843–867.

Perotti, R. (1996), “Growth, income distribution, and democracy: what the data say?”, *Journal of Economic Growth*, Vol. 1, pp. 149–187.

Quah, D. (1997), “Empirics for growth and distribution: stratification, polarization, and convergence clubs”, *Journal of Economic Growth*, Vol. 2, pp. 27–59.

Secretaria de Industria y Comercio: SIC (1970), *Ingresos y Egresos de la Población de México (1969–1970)*, Tomos I–IV, México.

Székely, M. (1995), “Economic liberalization, poverty and income distribution in Mexico”, Documento de trabajo de El Colegio de México, no. III-1995.

Further reading

INEGI, ENIGH surveys for 1984, 1989, 1992, 1994, 1996, 1998, 2000 and 2002.

INEGI, ENE surveys, 1991, 1993, 1998, and 2000.

Secretaria de Industria y Comercio: SIC (1960), *Ingresos y Egresos de la Población de México: Investigación por Muestreo Julio 1958*, México.

Secretaria de Programación y Presupuesto: SPP (1977), “Encuesta de ingresos y gastos de los hogares 1977”, primera observación gasto seminal y mensual, México.

This page intentionally left blank